Reverberant Signal Separation using Optimized Complex Sparse Nonnegative Tensor Deconvolution on Spectral Covariance Matrix

W.L. Woo[1,*], S.S. Dlay[1], Ahmed Al-Tmeme[2], B. Gao[3]

[1]School of Electrical and Electronic Engineering, Newcastle University, England, United Kingdom

[2]Information and Communication Eng. Dept., Al Khwarizmi Eng. College, University of Baghdad, Iraq

[3]School of Automation Engineering, University of Electronic Science and Technology of China, China

[*]Correspondence e-mail: lok.woo@ncl.ac.uk

Abstract — In this paper, an optimized complex nonnegative tensor factor 2D deconvolution (CNTF2D) is proposed to separate the sources that have been mixed in an underdetermined reverberant environment. Unlike conventional methods, the proposed model decomposition is performed directly on the statistics in the form of spectral covariance matrix instead of the data itself (i.e. the mixed signal). For faster convergence the model is adapted under the hybrid framework of the generalized expectation maximization and multiplicative update algorithms. This paper also proposes a solution to the issue of optimizing the model order i.e., number of components and convolutive parameters in the CNTF2D model. To this end, a latent-observation model based on Gamma-Exponential process is developed. In addition, the proposed Gamma-Exponential process can be used to initialize the parameterization of the CNTF2D. The proposed algorithm encodes a set of variable sparsity parameters derived from the Gibbs distribution. This permits a stable update and optimizes the CNTF2D with the correct degree of sparseness in the time-frequency domain. Experimental results on the underdetermined reverberant mixing environment have shown that the proposed algorithm is effective at separating the mixture with an average signal-to-distortion ratio of 2.5dB.

Index Terms **—** blind source separation, audio processing and analysis, spectral covariance, matrix factorization

1. INTRODUCTION

In source separation it is more realistic to consider the effect of the surrounding environment such as reflection of the sources. To address this issue, researchers have considered convolutive mixtures [1-7] instead of the instantaneous mixture [8-11]. However, the convolutive mixture is modeled under the narrowband approximation [4] that is not valid when the mixing filter length is greater than the Short-Time Fourier Transform (STFT) windows length, which is the case of the reverberant environment. Duong *et al.* [4] address this problem by considering the full rank spatial covariance matrix instead of the rank one. Arberet *et al.* [12] show that under the oracle initialization (where all the parameters are known) better results can be achieved if the nonnegative matrix factorization (NMF) is considered as a source variance as done by Duong *et al.* [4]. The NMF is too simplistic and does not efficiently model more complex sources such as polyphonic music. In addition, it is not always possible to adopt the oracle initialization approach. Furthermore, most NMF-based methods do not utilize the phase information of the channel. It was shown in [13] that incorporating the phase information into the NMF has the potential to increase the separation performance. In this paper, we propose a full rank Complex Nonnegative Tensor Factor 2D Deconvolution (CNTF2D) to model the spectral covariance matrix of the source image, taking into account the phase information and the model of the spatial covariance matrix. The Nonnegative Tensor Factorization (NTF) [14-17] has been previously shown to benefit from the complementary information in stereo channels. Contrary to NTF methods, the proposed CNTF2D will be optimized using the Generalized Expectation-Maximization and Multiplicative Update (GEM-MU) algorithm. It provides a probabilistic platform for joint estimation of the sources and the parameters as well as preserving the non-negativity constraints of the model. In addition, the GEM-MU algorithm accelerates the convergence speed of the parameters update. Concurrently, we allow variable sparsity to be encoded into the CNTF2D instead of using some heuristics approaches to fix them to a constant value. This variable sparsity will be developed based on the Gibbs distribution framework and optimized under the Itakura-Saito divergence. This will be contrasted with the uniform sparsity which assigns a fixed sparsity over all the temporal code of the factorization model [18]. Since acoustic sources such as speech change dynamically over time, uniform

sparsity will invariably lead to either over-sparseness (resulting in too many temporal code set to zero), or under-sparseness (too many ineffective temporal code). The proposed variable sparsity relieves this problem by optimizing the sparsity for each individual temporal code.

Furthermore, the issues of determining the required number of parameters in the model, that is, the number of components and convolutive parameters in the CNTF2D, as well as to initialize these parameters remain as challenges. A probabilistic method has been developed to meet these challenges. The Itakura-Saito (IS) divergence will be considered due to its scale invariant property [19]. Compared with the Least Square (LS) distance and Kullback-Leibler (KL) divergence cost functions, IS divergence deals with both low and high energy components with equal emphasis. Since both speech and music signals have large magnitude dynamic ranges, IS divergence provides a faithful measure between the observed data and the output generated from the adapted CNTF2D model. Furthermore, as each source has its own characteristics regarding the spectral and temporal features; such as the drum that has a high pitch with low temporal features and the opposite thing for the piano; then different convolutive parameters with different number of components are needed for each source. This variation in the number of components and convolutive parameters will be optimized using the variational Bayesian inference procedure. In addition, the proposed inference procedure will be used to initialize the CNTF2D model.

The novelty of this paper can be summarized as follows: Firstly, a complex NTF2D (CNTF2D) Gaussian model with full-rank spatial covariance matrix is developed to model the spectral covariance matrix of the source images in the STFT domain. Secondly, the parameters of the model are adapted using the hybrid GEM-MU algorithm for faster convergence and ensuring the preservation of non-negativity of the parameters. Thirdly, a variational Bayesian inference method is developed to estimate the number of components and number of convolutive parameters of the CNTF2D. Finally, to the best of authors' knowledge, the present work is the first to propose and investigate a CNTF2D for solving the underdetermined convolutive mixture separation instead of instantaneous mixture [20-24]. Furthermore, the proposed method is different from [21] that is also based on NMF2D, in that it considers the reverberations of the surrounding environment, it

considers both the temporal and pitch change of the sources through the NTF2D, and finally it considers the phase of the sources. The high level presentation of the proposed algorithm is shown in Fig. 1.

This paper is organized as follows: Section 2 is dedicated to the formulation of problem. The derivation of variable sparsity and the development of GEM-MU algorithm to work with the proposed CNTF2D model is presented in Section 3. In Section 4, the issue of model order estimation of the CNTF2D is considered. The Gamma-Exponential latent-observation model is used as a platform to develop a probabilistic framework for estimating the optimum model order for CNTF2D. Experimental results using the SiSEC'18 real datasets and comparison with a recent method will be presented in Section 5. Finally, Section 6 draws the conclusions.

2. PROBLEM FORMULATION

Let $x_i(t)$ be the observed multichannel signal that can be expressed in time domain as

$$x_i(t) = \sum_{j=1}^{J} c_{i,j}(t) + b_i(t), \qquad i = 1,2,\dots I\,,\ \ t = 1,\dots,T. \tag{1}$$

where $x_i(t) \in \mathbb{R}$ is the receiving signal from the *i*-th microphone (or channel), t and T are the time index and number of samples, respectively, $c_{i,j}(t) \in \mathbb{R}$ is the spatial image of the *j*-th source signal from the *i*-th microphone, J is the number of sources, I is the number of microphones, and $b_i(t) \in \mathbb{R}$ is some additive noise. The spatial image of the source $c_{i,j}(t)$ can be expressed as

$$c_{i,j}(t) = \sum_{l=0}^{L-1} a_{i,j}(l) s_j(t-l). \tag{2}$$

where $a_{i,j}(l) \in \mathbb{R}$ is the finite-impulse response of some (causal) filter, L is the filter length, and $s_j(t) \in \mathbb{R}$ is the *j*-th original source signal. By substituting eqn. (2) into eqn. (1), and assuming that the mixing channel is time-invariant, the STFT of (1) becomes

$$\boldsymbol{x}_{f,n} = \sum_{j=1}^{J} \boldsymbol{c}_{j,f,n} + \boldsymbol{b}_{f,n}. \tag{3}$$

where $\boldsymbol{x}_{f,n} = [x_{1,f,n} \quad \cdots \quad x_{I,f,n}]^H$, and $x_{i,f,n}$, $c_{i,j,f,n}$, $b_{i,f,n}$ are the complex-valued STFT of $x_i(t)$, $c_{i,j}(t)$, and $b_i(t)$, respectively. The term $f = 1, 2, \ldots, F$ is the frequency bin index, and $n = 1, 2, \ldots, N$ is the time frame index in the STFT. The spectral covariance matrix of $c_{i,j,f,n}$ (the complex-valued STFT of $c_{i,j}(t)$) defined as $\boldsymbol{\Sigma}_{j,f,n}^{(c)} = E\left[\boldsymbol{c}_{j,f,n}\boldsymbol{c}_{j,f,n}^H\right]$ where $E[\cdot]$ is the expectation can be expressed as

$$\boldsymbol{\Sigma}_{j,f,n}^{(c)} = \boldsymbol{\Sigma}_{j,f}^{(a)} v_{j,f,n}. \tag{4}$$

where $\boldsymbol{\Sigma}_{j,fn}^{(c)} \in \mathbb{C}^{I\times I}$, $\boldsymbol{\Sigma}_{j,f}^{(a)} \in \mathbb{C}^{I\times I}$ is the time-invariant spatial covariance matrix of the channel associated with the *j*-th source, $v_{j,f,n} \in \mathbb{R}$ is the *j*-th source variance in the spectrogram. The scalar representation of $\boldsymbol{\Sigma}_{j,f,n}^{(c)}$ is given by $\Sigma_{r,s,j,f,n}^{(c)}$ is the $(r,s)^{th}$ element of the $I \times I$ matrix $\boldsymbol{\Sigma}_{j,f,n}^{(c)}$. Similarly, $\Sigma_{r,s,j,f}^{(a)}$ is the $(r,s)^{th}$ element of the $I \times I$ matrix $\boldsymbol{\Sigma}_{j,f}^{(a)}$. For fixed r, s and j, it is noted that

(i) $\Sigma_{r,s,j,f}^{(a)}$ is a complex-valued scalar which can be expressed in terms of magnitude and phase:

$$\Sigma_{r,s,j,f}^{(a)} = \left|\Sigma_{r,s,j,f}^{(a)}\right| e^{\sqrt{-1}\,\alpha_f^{j,r,s}}. \tag{5}$$

where $\sqrt{-1}$ is adopted to represent the imaginary part.

(ii) $v_{j,f,n,}$ is a real-valued scalar which represents the source power spectrogram. Various models exist but for speech and audio signals, the NMF2D [24] is adopted:

$$v_{j,f,n} = \sum_{k=1}^{K_j} \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} g_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}. \tag{6}$$

where K_j is the number of components or frequency basis assigned to the *j*-th source, τ_{max} and ϕ_{max} refer to the number of temporal and frequency shifts in the model, $g_{f,k}^{\tau,j}$ represents the *k*-th spectral basis of the *j*-th source, and $h_{k,n}^{\phi,j}$ represents the *k*-th temporal code for each spectral basis element of the *j*-th source, for $f = 1, \ldots, F, n = 1, \ldots, N$, and $j = 1, \ \ldots, J$.

Using eqns. (5) and (6) in (4), we can write the latter as

$$\Sigma_{r,s,j,f,n}^{(c)} = \sum_{k=1}^{K_j} \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} w_{f-\phi,k}^{\tau,j,r,s} \, h_{k,n-\tau}^{\phi,j} \, e^{\sqrt{-1}\,\alpha_f^{j,r,s}} . \quad (7a)$$

where $w_{f,k}^{\tau,j,r,s} \triangleq \left|\Sigma_{r,s,j,f}^{(a)}\right| g_{f,k}^{\tau,j}$ is the combined channel-source spectral basis. For the case where the channels are time-varying, $\Sigma_{j,f}^{(a)}$ has a dependency on time frame n and this representation can be absorbed into the temporal code and phase spectrum. Hence (7a) can be generalized to

$$\Sigma_{r,s,j,f,n}^{(c)} = \sum_{k=1}^{K_j} \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} w_{f-\phi,k}^{\tau,j,r,s} \, h_{k,n-\tau}^{\phi,j,r,s} \, e^{\sqrt{-1}\,\alpha_{f,n}^{j,r,s}} . \quad (7b)$$

where $\alpha_{f,n}^{j,r,s}$ is equivalent to the time-varying phase spectrum [25]. The dimension of the variables are as follows: $\Sigma_{r,s,j,f,n}^{(c)} \in \mathbb{C}^{F\times N\times J\times I\times I}$, $w_{f,k}^{\tau,j,r,s} \in \mathbb{R}^{F\times K\times \tau_{max}\times I\times I}$, $h_{k,n-\tau}^{\phi,j,r,s} \in \mathbb{R}^{K\times N\times \phi_{max}\times I\times I}$ and $\alpha_{f,n}^{j,r,s} \in \mathbb{R}^{F\times N\times J\times I\times I}$. The spectral covariance matrix of $x_{f,n}$ can be expressed as $\boldsymbol{\Sigma}_{f,n}^{(x)} = E\left[\boldsymbol{x}_{f,n}\boldsymbol{x}_{f,n}^H\right] = \sum_{j=1}^{J} \boldsymbol{\Sigma}_{j,f,n}^{(c)} + \boldsymbol{\Sigma}_f^{(b)}$ where $\boldsymbol{\Sigma}_f^{(b)}$ is the time invariant noise covariance matrix. Its scalar form can be expressed as

$$\begin{aligned}\Sigma_{r,s,f,n}^{(x)} &= \sum_{j=1}^{J} \Sigma_{r,s,j,f,n}^{(c)} + \Sigma_{r,s,f}^{(b)} \\ &= \sum_{j=1}^{J} \sum_{k=1}^{K_j} \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} w_{f-\phi,k}^{\tau,j,r,s} \, h_{k,n-\tau}^{\phi,j,r,s} \, e^{\sqrt{-1}\,\alpha_{f,n}^{j,r,s}} + \Sigma_{r,s,f}^{(b)}. \end{aligned} \quad (8)$$

Most conventional source separation methods work on the spectrogram of the data samples; however, the proposed method performs complex matrix factorization on the spectral covariance matrices where the latter has to be constructed by computing the first and second order statistics of the data spectrogram as shown in Section 3.2. Thus, a point of departure between the proposed method and other conventional algorithms is that the former works directly on the statistics (i.e., spectral covariance matrices) instead on the data samples (i.e., the time-domain mixture signal or its spectrogram) [21, 24].

3. PROPOSED ESTIMATION ALGORITHM

In this section, the source model and the Generalized Expectation-Maximization with Multiplicative Update (GEM-MU) algorithm will be developed. The GEM-MU algorithm is formulated in two steps, namely, E-step and M-step. To pave the way forward for the estimation of the parameters, a graphical model of the proposed CNTF2D has been constructed. The performance of matrix factorization depends considerably on the sparsity of the solution. Thus a sub-section is dedicated on the development of adaptive estimation of the sparsity for the temporal codes. Finally, it is shown how the separated image sources are reconstructed using the minimum mean square error estimate.

3.1. Source model

The spatial image of the sources can be modeled as realization of zero-mean proper complex distribution

$$\boldsymbol{c}_{j,f,n} \sim \mathcal{N}_c\left(0, \boldsymbol{\Sigma}_{j,f,n}^{(c)}\right). \tag{9}$$

where $\mathcal{N}_c(\mu, \Sigma)$ is proper complex Gaussian distribution [26] and its probability density function (pdf) can be expressed as

$$\mathcal{N}_c\left(0, \boldsymbol{\Sigma}_{j,f,n}^{(c)}\right) \triangleq \frac{1}{det\left(\pi \boldsymbol{\Sigma}_{j,f,n}^{(c)}\right)} e^{-\left(\boldsymbol{c}_{j,f,n}^{H} \boldsymbol{\Sigma}_{j,f,n}^{(c)^{-1}} \boldsymbol{c}_{j,f,n}\right)}. \tag{10}$$

By substituting eqn. (7) into eqn. (9) we have

$$\boldsymbol{c}_{j,f,n} \sim \mathcal{N}_c\left(\mathbf{0}, \left[\sum_{k,\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s}\, h_{k,n-\tau}^{\phi,j,r,s}\, e^{\sqrt{-1}\,\alpha_{f,n}^{j,r,s}}\right]_{r,s}\right). \tag{11}$$

which is a zero mean with complex covariance matrix whose $(r,s)^{th}$ element is given by $\sum_{k,\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s}\, h_{k,n-\tau}^{\phi,j,r,s}\, e^{\sqrt{-1}\,\alpha_{f,n}^{j,r,s}}$ The noise $\boldsymbol{b}_{f,n}$ in eqn. (3) is assumed to be time invariant, stationary and spatially uncorrelated, i.e.

$$\boldsymbol{b}_{f,n} \sim \mathcal{N}_c\left(0, \boldsymbol{\Sigma}_{f}^{(b)}\right). \tag{12}$$

and its distribution can be expressed as

$$\mathcal{N}_c\left(0, \boldsymbol{\Sigma}_f^{(\boldsymbol{b})}\right) \triangleq \frac{1}{det\left(\pi \boldsymbol{\Sigma}_f^{(\boldsymbol{b})}\right)} e^{-\left(\boldsymbol{b}_{f,n}^H \boldsymbol{\Sigma}_f^{(\boldsymbol{b})^{-1}} \boldsymbol{b}_{f,n}\right)}. \quad (13)$$

3.2. Generalized Expectation-Maximization with Multiplicative Update (GEM-MU) algorithm

The source images, noise and their spectral covariances will be estimated using the GEM algorithm while $W = \left\{w_{f,k}^{\tau,j,r,s}\right\}$, $H = \left\{ h_{k,n}^{\phi,j,r,s}\right\}$, and $\alpha = \left\{\alpha_{f,n}^{j,r,s}\right\}$ will be estimated in the M step using the MU algorithm. The model parameters are $\Theta = \{W, H, \Sigma^{(b)}, \Lambda, \alpha\}$, with observations $X = \{\boldsymbol{x}_{f,n}\}$. $\Lambda = \left\{\lambda_{k,n}^{\phi,j,r,s}\right\}$ is a tensor that contains the sparsity terms, that are added to the cost function as a constraint in order to ensure that only a few units (out of a large population) in the temporal code will be active at the same time [27]. We define $C = \{\boldsymbol{c}_{j,f,n}\}$ and $\Sigma^{(b)} = \left\{\boldsymbol{\Sigma}_f^{(\boldsymbol{b})}\right\}$ according to (11) and (13), respectively. To pave the way forward for the estimation of the parameters, a graphical model of the proposed CNTF2D has been constructed and is shown in Fig. 2.The nodes represent random variables (shaded node refers to observed variable and unshaded node refers to latent variable) and dots represent parameters. Firstly, we determine the posterior distribution of C, W, H:

$$P\left(C, W, H \middle| X, \Sigma^{(b)}, \Lambda, \alpha\right) = \frac{P\left(X \middle| C, W, H, \Sigma^{(b)}, \Lambda, \alpha\right) P\left(C, W, H, \Sigma^{(b)}, \Lambda, \alpha\right)}{P\left(X, \Sigma^{(b)}, \Lambda, \alpha\right)}.$$

From the graphical model, it can be deduced that

(i) $P\left(X \middle| C, W, H, \Sigma^{(b)}, \Lambda, \alpha\right) = P(X|C, \Theta)$ with $\Theta = \{W, H, \Sigma^{(b)}, \Lambda, \alpha\}$

(ii) $P\left(C, W, H, \Sigma^{(b)}, \Lambda, \alpha\right) = P\left(C \middle| W, H, \Sigma^{(b)}, \Lambda, \alpha\right) P\left(W, H, \Sigma^{(b)}, \Lambda, \alpha\right)$ where $P\left(C \middle| W, H, \Sigma^{(b)}, \Lambda, \alpha\right) = P(C|W, H, \alpha)$ and $P\left(W, H, \Sigma^{(b)}, \Lambda, \alpha\right) = P(W, H|\Lambda) P\left(\Sigma^{(b)}, \Lambda, \alpha\right)$

(iii) $P\left(X, \Sigma^{(b)}, \Lambda, \alpha\right) = P\left(X \middle| \Sigma^{(b)}, \Lambda, \alpha\right) P\left(\Sigma^{(b)}, \Lambda, \alpha\right)$

Therefore,

$$P\left(C,W,H\middle|X,\Sigma^{(b)},\Lambda,\alpha\right)=\frac{P(X|C,\Theta)P(C|W,H,\alpha)P(W,H|\Lambda)P\left(\Sigma^{(b)},\Lambda,\alpha\right)}{P\left(X\middle|\Sigma^{(b)},\Lambda,\alpha\right)P(\Sigma^{(b)},\Lambda,\alpha)}$$

$$=\frac{P(X|C,\Theta)P(C|W,H,\alpha)P(W,H|\Lambda)}{P\left(X\middle|\Sigma^{(b)},\Lambda,\alpha\right)}. \quad (14)$$

From the graphical model, it can also be further deduced that $P(X|C,\Theta)=P\left(X\middle|C,\Sigma^{(b)}\right)$ and $P(W,H|\Lambda)=P(W)P(H|\Lambda)$. From eqn. (14), the negative log-posterior is given by

$$-\log P\left(C,W,H\middle|X,\Sigma^{(b)},\Lambda,\alpha\right)=-\log P(X|C,\Theta)-\log P(C|W,H,\alpha)-\log P(W,H|\Lambda)+const. \quad (15)$$

where $-\log P\left(X\middle|\Sigma^{(b)},\Lambda,\alpha\right)$ is treated as a normalizing constant. In eqn. (15), the first term on the right hand side corresponds to the data log-likehood, the second term is the log-likehood of the spatial source images given the CNTF2D parameters, and the third term is the log-likehood of the channel-source spectral basis and temporal code. One can think that the incorporation of the second and third terms into the data log-likehood serve as a form of probabilistic regularization and allows the user to add prior information into the solution. The log-posterior probability will be computed by the GEM-MU based variable sparsity CNTF2D in the following sections.

3.2.1. E-Step: Conditional expectations of natural statistics

In the E-step, we determine the conditional expectations of the natural statistics. The log-likelihood in eqn. (15) is given by

$$\log P(X|C,\Theta)=\sum_{f,n}tr\left(\boldsymbol{\Sigma}_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}\boldsymbol{x}_{f,n}^{H}\right)+\log\left|\pi\boldsymbol{\Sigma}_{f,n}^{(x)}\right|. \quad (16)$$

where $tr(.)$ refers to the trace operator. The conditional expectation of the natural statistics $\widehat{\boldsymbol{R}}_{j,f,n}^{(c)},\widehat{\boldsymbol{R}}_{f}^{(b)}$, $\widehat{\boldsymbol{\Sigma}}_{j,f,n}^{(c)},\widehat{\boldsymbol{\Sigma}}_{f}^{(b)},\hat{\boldsymbol{c}}_{j,f,n}$ and $\widehat{\boldsymbol{b}}_{f,n}$ can be obtained using the complete data likelihood $\log P(X,C|\Theta)$ as follows:

$$\widehat{\boldsymbol{R}}_{j,f,n}^{(c)}=\hat{\boldsymbol{c}}_{j,f,n}\hat{\boldsymbol{c}}_{j,f,n}^{H}+\widehat{\boldsymbol{\Sigma}}_{j,f,n}^{(c)}. \quad (17)$$

$$\widehat{\boldsymbol{\Sigma}}_{j,f,n}^{(c)} = \left(\boldsymbol{I} - \boldsymbol{\Sigma}_{j,f,n}^{(c)} \boldsymbol{\Sigma}_{f,n}^{(x)^{-1}}\right) \boldsymbol{\Sigma}_{j,f,n}^{(c)} . \tag{18}$$

$$\hat{\boldsymbol{c}}_{j,f,n} = \boldsymbol{\Sigma}_{j,f,n}^{(c)} \boldsymbol{\Sigma}_{f,n}^{(x)^{-1}} \boldsymbol{x}_{f,n}. \tag{19}$$

$$\widehat{\boldsymbol{R}}_{f}^{(b)} = \widehat{\boldsymbol{b}}_{f,n} \widehat{\boldsymbol{b}}_{f,n}^{H} + \widehat{\boldsymbol{\Sigma}}_{f}^{(b)}. \tag{20}$$

$$\widehat{\boldsymbol{\Sigma}}_{f}^{(b)} = \left(\boldsymbol{I} - \boldsymbol{\Sigma}_{f}^{(b)} \boldsymbol{\Sigma}_{f,n}^{(x)^{-1}}\right) \boldsymbol{\Sigma}_{f}^{(b)}. \tag{21}$$

$$\widehat{\boldsymbol{b}}_{f,n} = \boldsymbol{\Sigma}_{f}^{(b)} \boldsymbol{\Sigma}_{f,n}^{(x)^{-1}} \boldsymbol{x}_{f,n}. \tag{22}$$

The derivation for the above expressions follows from the linear complex Gaussian model of eqn. (3) in the STFT domain.

3.2.2. M-Step: Update of parameters

In the M-step, the parameters of the model are updated based on the conditional expectations obtained from the natural statistics in eqns. (17)-(22). The scalar form of $\hat{R}_{j,f,n}^{(c)}$ can be expressed as follows $\hat{R}_{r,s,,j,f,n}^{(c)} = \left\{\hat{R}_{j,f,n}^{(c)}\right\}_{r,s}$ which is the$(r,s)^{th}$ element of the $I \times I$ matrix $\hat{R}_{j,f,n}^{(c)}$. The second term in the right hand side of (15) can be expressed with IS divergence [15] as

$$-\log P(C|W,H,\alpha) = \sum_{r,s,j,f,n} D_{IS}\left(\hat{R}_{r,s,j,f,n}^{(c)} \middle| \Sigma_{r,s,j,f,n}^{(c)}\right). \tag{23}$$

The third term in the right hand side of (15) is the prior information on W and H. An improper prior is assumed for W and factor-wise normalized to unit length i.e. $p(W) = \prod_j \delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right)$ where $\boldsymbol{W}^j = \left\{w_{f-\phi,k}^{\tau,j,r,s}\right\}$ is the spectral basis that belongs to the *j*-th source. Each element of *H* has independent decay parameter $\lambda_{k,n}^{\phi,j,r,s}$ with exponential distribution:

$$\begin{aligned}\log p(W,H|\Lambda) &= \log \prod_j \delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right) + \log\left(\prod_{j,k} p\left(H_k^j|\Lambda_k^j\right)\right) \\ &= \sum_j \log \delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right) - \sum_{r,s,j,k,n\phi,} \left(\lambda_{k,n}^{\phi,j,r,s} h_{k,n}^{\phi,j,r,s} - \log \lambda_{k,n}^{\phi,j,r,s}\right).\end{aligned} \tag{24}$$

Inserting eqns. (16), (23), (24) to (15) yields the following:

$$
\begin{aligned}
&-\log P\left(C,W,H\middle|X,\Sigma^{(b)},\Lambda,\alpha\right) \\
&\quad=-\sum_{f,n} tr\left(\boldsymbol{\Sigma}_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}\boldsymbol{x}_{f,n}^{H}\right)-\log\left|\pi\boldsymbol{\Sigma}_{f,n}^{(x)}\right| \\
&\quad+\sum_{r,s,j,k,f,n}\left(\hat{R}_{r,s,j,f,n}^{(c)}\Sigma_{r,s,j,f,n}^{(c)^{-1}}-\log\left(\hat{R}_{r,s,j,f,n}^{(c)}\Sigma_{r,s,j,f,n}^{(c)^{-1}}\right)-1\right)-\sum_{j}\log\delta\left(\left\|\boldsymbol{W}^{j}\right\|_{2}-1\right) \\
&\quad+\sum_{r,s,j,k,n,\phi}\lambda_{k,n}^{\phi,j,r,s}h_{k,n}^{\phi,j,r,s}-\sum_{r,s,j,k,n,\phi}\log\lambda_{k,n}^{\phi,j,r,s}. \qquad (25)
\end{aligned}
$$

The differentiation of eqn. (25) with respect to $w_{f-\phi,k}^{\tau,j,r,s}$, $h_{k,n-\tau}^{\phi,j,r,s}$, and $\alpha_{f,n}^{j,r,s}$ gives the followings:

$$
\begin{aligned}
&\frac{\partial}{\partial w_{f,k}^{\tau,j,r,s}}\log P\left(C,W,H\middle|X,\Sigma^{(b)},\Lambda,\alpha\right) \\
&\quad=-\sum_{n,\phi}\hat{R}_{r,s,j,f+\phi,n}^{(c)}\Sigma_{r,s,j,f+\phi,n}^{(c)^{-2}}e^{-\sqrt{-1}\,\alpha_{f+\phi,n}^{j,r,s}}h_{k,n-\tau}^{\phi,j,r,s}+\sum_{\phi,n}\Sigma_{r,s,j,f+\phi,n}^{(c)^{-1}}e^{-\sqrt{-1}\,\alpha_{f+\phi,n}^{j,r,s}}h_{k,n-\tau}^{\phi,j,r,s}. \qquad (26)
\end{aligned}
$$

Likewise,

$$
\begin{aligned}
&\frac{\partial}{\partial h_{k,n}^{\phi,j,r,s}}\log P\left(C,W,H\middle|X,\Sigma^{(b)},\Lambda,\alpha\right) \\
&\quad=-\sum_{f\tau}\hat{R}_{r,s,j,f,n+\tau}^{(c)}\Sigma_{r,s,j,f,n+\tau}^{(c)^{-2}}\,e^{-\sqrt{-1}\,\alpha_{f,n+\tau}^{j,r,s}}w_{f-\phi,k}^{\tau,j,r,s}+\sum_{f\tau}\Sigma_{r,s,j,f,n+\tau}^{(c)^{-1}}e^{-\sqrt{-1}\,\alpha_{f,n+\tau}^{j,r,s}}\,w_{f-\phi,k}^{\tau,j,r,s}+\lambda_{k,n}^{\phi,j,r,s}. \qquad (27)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\frac{\partial}{\partial \alpha_{f,n}^{j,r,s}}\log P\left(C,W,H\middle|X,\Sigma^{(b)},\Lambda,\alpha\right) \\
&\quad=-\sqrt{-1}\hat{R}_{r,s,j,f,n}^{(c)}\Sigma_{r,s,j,f,n}^{(c)^{-2}}e^{-\sqrt{-1}\,\alpha_{f,n}^{j,r,s}}\sum_{\tau,\phi,k}w_{f-\phi,k}^{\tau,j,r,s}\,h_{k,n-\tau}^{\phi,j,r,s}+\sqrt{-1}\Sigma_{r,s,j,f,n}^{(c)^{-1}}e^{-\sqrt{-1}\,\alpha_{f,n}^{j,r,s}}\sum_{\tau,\phi,k}w_{f-\phi,k}^{\tau,j,r,s}h_{k,n-\tau}^{\phi,j,r,s}. \qquad (28)
\end{aligned}
$$

Therefore, the MU rules for $w_{f-\phi,k}^{\tau,j,r,s}$, $h_{k,n-\tau}^{\phi,j,r,s}$, and $\alpha_{f,n}^{j,r,s}$ can be respectively formulated as

$$w_{f,k}^{\tau,j,r,s} \leftarrow w_{f,k}^{\tau,j,r,s} \left(\frac{\sum_{n,\phi} \hat{R}_{r,s,j,f+\phi,n}^{(c)} \Sigma_{r,s,j,f+\phi,n}^{(c)^{-2}} e^{-\sqrt{-1}\,\alpha_{f+\phi,n}^{j,r,s}} h_{k,n-\tau}^{\phi,j,r,s}}{\sum_{\phi,n} \Sigma_{r,s,j,f+\phi,n}^{(c)^{-1}} e^{-\sqrt{-1}\,\alpha_{f+\phi,n}^{j,r,s}} h_{k,n-\tau}^{\phi,j,r,s}} \right). \tag{29}$$

$$h_{k,n}^{\phi,j,r,s} \leftarrow h_{k,n}^{\phi,j,r,s} \left(\frac{\sum_{f\tau} \hat{R}_{r,s,j,f,n+\tau}^{(c)} \Sigma_{r,s,j,f,n+\tau}^{(c)^{-2}} e^{-\sqrt{-1}\,\alpha_{f,n+\tau}^{j,r,s}} w_{f-\phi,k}^{\tau,j,r,s}}{\sum_{f\tau} \Sigma_{r,s,j,f,n+\tau}^{(c)^{-1}} e^{-\sqrt{-1}\,\alpha_{f,n+\tau}^{j,r,s}} w_{f-\phi,k}^{\tau,j,r,s} + \lambda_{k,n}^{\phi,j,r,s}} \right). \tag{30}$$

$$e^{\sqrt{-1}\,\alpha_{f,n}^{j,r,s}} \leftarrow \frac{\hat{R}_{r,s,j,f,n}^{(c)}}{\sum_{\tau,\phi,k} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}}. \tag{31}$$

In eqn. (29), in order to satisfy the constraint $\delta\left(\left\|\boldsymbol{W}^{j}\right\|_{2}-1\right)$, each spectral basis is explicitly normalized to unity i.e. $w_{f,k}^{\tau,j,r,s} = w_{f,k}^{\tau,j,r,s} \Big/ \sqrt{\sum_{f,\tau,k} \left(w_{f,k}^{\tau,j,r,s}\right)^{2}}$.

3.2.3. Estimation of variable sparsity using Gibbs distribution

For the sparsity term, the update is obtained by maximizing the log-posterior as follows:

$$\hat{\lambda} = \arg\max_{\lambda} \log P\left(C, W, H \,\middle|\, X, \Sigma^{(b)}, \Lambda, \alpha\right). \tag{32}$$

Solving $\frac{\partial}{\partial \lambda} \log P\left(C, W, H \,\middle|\, X, \Sigma^{(b)}, \Lambda, \alpha\right) = 0$ will lead to

$$\hat{\lambda}_{k,n}^{\phi,j,r,s} = \frac{1}{h_{k,n}^{\phi,j,r,s}} \quad (\text{or in matrix form } \Lambda = 1 \cdot / H). \tag{33}$$

where *"·/"* represents element-wise division. Since we are seeking a sparse H, then the above solution (33) will yield divergent updates in cases where $h_{k,n}^{\phi,j,r,s} = 0$. Therefore, a better approximation to account for variability of H is required. We partition H into two distinct subsets of positive values $h_{k,n}^{\phi,j,r,s} > 0$ and zero

value $h_{k,n}^{\phi,j,r,s} = 0$, and develop a probability distribution for each subset. For any distribution $Q(\underline{h})$, the log-likelihood function satisfies the Jensen's inequality:

$$\log P(X|\underline{\lambda}) \geq \int Q(\underline{h}) \log\left(\frac{P(X,\underline{h}|\underline{\lambda})}{Q(\underline{h})}\right) d\underline{h}. \tag{34}$$

In eqn. (34), both H and Λ are vectorized into column vectors as $\underline{h}$ and $\underline{\lambda}$ which have dimension $D \times 1$ where $D = K \times N \times \Phi_{max} \times I^2$. The elements of $\underline{h}$ and $\underline{\lambda}$ are denoted as h_p and λ_p, respectively, for $p = 1, 2, \ldots, D$. By substituting eqn. (34) into eqn. (32), this leads to

$$\underline{\hat{\lambda}} = arg\, \max_{\underline{\lambda}} \int Q(\underline{h}) \left(log\, \lambda_p - \lambda_p h_p\right) d\underline{h}. \tag{35}$$

Eqn. (35) can be solved as follows:

$$\frac{\partial}{\partial \lambda_p} \int Q(\underline{h}) \left(\log \lambda_p - \lambda_p\, h_p\right) d\underline{h} = 0$$

Therefore,

$$\hat{\lambda}_p = \frac{1}{\int h_p Q(\underline{h}) d\underline{h}} = \frac{1}{E_{Q(\underline{h})}[h_P]}. \tag{36}$$

where $E_{Q(\underline{h})}[h_P]$ is the expectation of h_P under the distribution $Q(\underline{h})$. However, eqn. (36) cannot be solved analytically therefore we will approximate $Q(\underline{h})$ with respect to the mode of distribution h_p. As h_p can be partitioned into distinct subsets of positive value $(\underline{h_M})\ \forall_m \in M$ such that $h_m > 0$, and zero value $(\underline{h_L})\ \forall_l \in L$ such that $h_l = 0$. It then follows from eqn. (25) and by using the reverse Triangle Inequality [28], for any h_p, h_m and h_l satisfying the above, it can be shown that:

$$F(\underline{h}) \equiv \sum_{r,s,j,f,p} D_{IS}\left(\hat{R}_{r,s,j,f,p}^{(c)} \middle| \Sigma_{r,s,j,f,p}^{(c)}\right) + \sum_{p} \left(\lambda_p h_p - \log \lambda_p\right)$$

$$\geq \sum_{r,s,j,f,m} D_{IS}\left(\hat{R}^{(c)}_{r,s,j,f,m}\middle|\, \Sigma^{(c)}_{r,s,j,f,m}\right) + \sum_{m}(\lambda_m h_m - \log\lambda_m)$$

$$+ \sum_{r,s,j,f,l} D_{IS}\left(\hat{R}^{(c)}_{r,s,j,f,l}\middle|\, \Sigma^{(c)}_{r,s,j,f,l}\right) + \sum_{l}(\lambda_l h_l - \log\lambda_l)$$

Thus,

$$F(\underline{h}) \geq F(\underline{h}_L) + F(\underline{h}_M). \tag{37}$$

In this paper, we will use the Gibbs distribution as the approximate distribution $Q(\underline{h})$ i.e. $Q(\underline{h}) = Z_h^{-1}\, exp[-F(\underline{h})]$ where $Z_h = \int exp[-F(\underline{h})]d\underline{h}$ therefore $Q(\underline{h})$ can be factorized into a product of $Q_L(\underline{h}_L)$ and $Q_M(\underline{h}_M)$:

$$Q(\underline{h}) \approx Z_h^{-1}\, exp[-F(\underline{h}_L) - F(\underline{h}_M)]$$

$$= \frac{1}{Z_L} exp[-F(\underline{\boldsymbol{h}}_L)]\frac{1}{Z_M} exp[-F(\underline{\boldsymbol{h}}_M)]$$

$$= Q_L(\underline{h}_L)Q_M(\underline{h}_M). \tag{38}$$

where $Z_L = \int exp[-F(\underline{\boldsymbol{h}}_L)]\, d\underline{\boldsymbol{h}}_L$ and $Z_M = \int exp[-F(\underline{\boldsymbol{h}}_M)]d\underline{\boldsymbol{h}}_M$. This leads to $E_{Q_M(\underline{h}_M)}[h_P] = h_P$ (which is optimized in eqn. (30)), and $E_{Q_L(\underline{h}_L)}[h_P] = u_l$ where u_l is the variational parameter. Therefore, eqn. (36) is given by

$$\hat{\lambda}_p = \begin{cases} 1/h_p & \forall_p \in M \\ 1/u_p & \forall_p \in L \end{cases}. \tag{39}$$

where

$$u_p \leftarrow u_p \frac{-b_p + \sqrt{b_p^2 + 4\dfrac{(\tilde{\Theta}\underline{u})_p}{u_p}}}{2(\tilde{\Theta}\underline{u})_p}. \tag{40}$$

$$\tilde{\Theta} = \mathrm{diag}(\Theta_p). \tag{41a}$$

$$\Theta_p = \sum_{r,s,j,k,f,\phi} \left(-2\left(w_{f-\phi,k}^{\tau,j,r,s}\right)^2 e^{-2\sqrt{-1}\,\alpha_{f,p}^{j,r,s}} \hat{R}_{r,s,j,f,p}^{(c)} \Sigma_{r,s,j,f,p}^{(c)^{-3}} + \left(w_{f-\phi,k}^{j}\right)^2 e^{-2\sqrt{-1}\alpha_{f,p}^{j,r,s}} \Sigma_{r,s,j,f,p}^{(c)^{-2}}\right). \quad (41b)$$

$$b_p = \sum_{r,s,j,k,f,\phi} \left(\hat{R}_{r,s,j,f,p}^{(c)} \Sigma_{r,s,j,f,p}^{(c)^{-2}} e^{-\sqrt{-1}\,\alpha_{f,p}^{j,r,s}} w_{f-\phi,k}^{\tau,j,r,s} - \Sigma_{r,s,j,f,p}^{(c)^{-1}} e^{-\sqrt{-1}\,\alpha_{f,p}^{j,r,s}} w_{f-\phi,k}^{\tau,j,r,s} - \lambda_p\right). \quad (42)$$

The detailed derivation of the variational parameter u_p can be found in Appendix A1.

3.2.4. Components Reconstruction

The estimated STFT source spatial image $\hat{c}_{j,f,n}$ can be reconstructed by using the multichannel Wiener filter that obtained by the minimum mean square error (MMSE) estimate $\hat{\boldsymbol{c}}_{j,f,n} = \mathbb{E}\left[\boldsymbol{c}_{j,f,n} \middle| \boldsymbol{x}_{f,n}; \Theta\right]$as in eqn. (19). The multichannel Wiener filter takes all the source spatial image components instead of the dominant one, as in the binary masking. Due to the linearity of the STFT, the inverse-STFT (with dual synthesis window [29]) can be used to transform the source spatial image to time domain.

4. MODEL ORDER ESTIMATION

In this section, the issue of model order estimation of the CNTF2D is considered. This includes estimation of the number of effective components and the number of convolutive parameters (which refers to the number of τ components and number of ϕ components of the CNTF2D model). It is also shown how spectral and temporal tensors of the CNTF2D can be initialized.

4.1. Latent-observation model

The Gamma-Exponential process is proposed to estimate the convolutive parameters and the number of components of the CNTF2D model. In Section 3, $\boldsymbol{W}$ is set as improper prior and $\boldsymbol{H}$ as generalized exponential. For the purpose of model order estimation, we generalize the previous setting to generative distributions:

$$w_{f,k}^{\tau,j,r,s} \sim Gamma\left(a_k^{\tau,j,r,s}, a_k^{\tau,j,r,s}\right). \quad (43)$$

$$h_{k,n}^{\phi,j,r,s} \sim Gamma\left(b_k^{\phi,j,r,s}, b_k^{\phi,j,r,s}\right). \tag{44}$$

where first parameter represents the shape and the second parameter is the rate. The magnitude of the spectral covariance matrix of the mixture signal is modelled as exponential distribution. We will also introduce a hidden tensor of nonnegative values $\theta_k^{\tau,\phi}$ that weight each element of the factor model i.e. $\sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}$ such that the number of components and convolutive parameters are inferred automatically based on the observed mixture data. In this way, the proposed model will retain a finite number of each subset corresponding to the active elements in θ. Using the above, the spectral covariance matrix of the mixture signal is given by

$$\left|\Sigma_{r,s,f,n}^{(x)}\right| \sim Exp\left(\sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}\right). \tag{45}$$

and

$$\theta_k^{\tau,\phi} \sim Gamma\left(\frac{\vartheta_{\tau,\phi}}{L_\theta}, \vartheta_{\tau,\phi} c\right). \tag{46}$$

where *Gamma* and *Exp* denote the Gamma and exponential distributions, respectively, $L_\theta = L + \phi_{max} + \tau_{max}$ with L being a positive number. It should be noted that L_θ defines the truncation level and it increases to infinity, then $\left\{\theta_k^{\tau,\phi}\right\}$ approximates an infinite sequence drawn from a gamma process with shape parameter $\vartheta_{\tau,\phi}$ and inverse-scale parameter $\vartheta_{\tau,\phi} c$. A property of this consequence is that the number of elements K greater than some number $\epsilon > 0$ is finite almost surely [30]. Specifically, we have $K \sim Poisson\left(\frac{1}{c}\int_\epsilon^\infty x^{-1} e^{-x\vartheta_{\tau,\phi} c}\, dx\right)$. For truncation levels L_θ that are sufficiently large relative to the shape parameter $\vartheta_{\tau,\phi}$, we would likewise expect that only a few of the L_θ elements of $\theta_k^{\tau,\phi}$ will be substantially greater than *0*. The expected value of $\Sigma_{r,s,f,n}^{(x)}$ under this model is constant with respect to L_θ, $\vartheta_{\tau,\phi}$, $a_k^{\tau,j,r,s}$ and $b_k^{\phi,j,r,s}$:

$$\mathbb{E}_p\left[\left|\Sigma^{(x)}_{r,s,f,n}\right|\right] = \sum\nolimits_{j,k,\tau,\phi} \mathbb{E}_p\left[\theta^{\tau,\phi}_k\right] \mathbb{E}_p\left[w^{\tau,j,r,s}_{f-\phi,k}\right] \mathbb{E}_p\left[h^{\phi,j,r,s}_{k,n-\tau}\right] \quad = \frac{1}{c}. \tag{47}$$

Eqn. (47) suggests setting the expected mean of the spatial covariance matrix under the prior equal to its empirical mean $\hat{\Sigma}^{(x)}_{r,s,f,n}$ by setting $c = 1/\hat{\Sigma}^{(x)}_{r,s,f,n}$. In this paper, we use the Generalized Inverse-Gaussian (GIG) [31] family to approximate the posterior distribution. The GIG for our model is given by:

$$q\left(w^{\tau,j,r,s}_{f,k}\right) = GIG\left(\zeta^{\tau,r,s}_{w,f,k}, \psi^{\tau,r,s}_{w,f,k}, \beta^{\tau,r,s}_{w,f,k}\right). \tag{48}$$

$$q\left(h^{\phi,j,r,s}_{k,n}\right) = GIG\left(\zeta^{\phi,r,s}_{h,k,n}, \psi^{\phi,r,s}_{h,k,n}, \beta^{\phi,r,s}_{h,k,n}\right). \tag{49}$$

$$q\left(\theta^{\tau,\phi}_{k}\right) = GIG\left(\zeta^{\tau,\phi}_{\theta,k}, \psi^{\tau,\phi}_{\theta,k}, \beta^{\tau,\phi}_{\theta,k}\right). \tag{50}$$

where

$$GIG(y;\zeta,\psi,\beta) = \frac{(\psi/\beta)^{\zeta/2}}{2\mathcal{K}_\zeta\left(2\sqrt{\beta\psi}\right)}\, y^{\zeta-1} exp(-(\beta y^{-1} + \psi y)/2). \tag{51}$$

for $y \geq 0$, $\zeta \geq 0$ and $\beta \geq 0$, and $\mathcal{K}_\zeta(\cdot)$ is the modified Bessel function of the third kind with index ζ. The expectation under q can be computed for any variable $y \sim GIG(\zeta,\psi,\beta)$:

$$\mathbb{E}_q[y] = \frac{\sqrt{\beta/\psi}\mathcal{K}_{\zeta+1}\left(2\sqrt{\psi\beta}\right)}{\mathcal{K}_\zeta\left(2\sqrt{\psi\beta}\right)}. \tag{52}$$

$$\mathbb{E}_q[1/y] = \frac{\sqrt{\psi/\beta}\mathcal{K}_{\zeta-1}\left(2\sqrt{\psi\beta}\right)}{\mathcal{K}_\zeta\left(2\sqrt{\psi\beta}\right)}. \tag{53}$$

By using the Jensen's inequality, the posterior distribution of $\left|\Sigma^{(x)}_{r,s,f,n}\right|$ is bounded below as

$$p\left(\left|\Sigma_{r,s,f,n}^{(x)}\right| \left|\vartheta_{\tau,\phi}, a_k^{\tau,j,r,s}, b_k^{\phi,j,r,s}, c\right.\right)$$

$$\geq \mathbb{E}_q\left[log\, p\left(\left|\Sigma_{r,s,f,n}^{(x)}\right| \left|\left\{w_{f,k}^{\tau,j,r,s}\right\}, \left\{h_{k,n}^{\phi,j,r,s}\right\}, \left\{\theta_k^{\tau,\phi}\right\}\right.\right)\right] + \mathbb{E}_q\left[log\, p\left(\left\{w_{f,k}^{\tau,j,r,s}\right\} \middle| a_k^{\tau,j,r,s}\right)\right]$$

$$- \mathbb{E}_q\left[log\, q\left(\left\{w_{f,k}^{\tau,j,r,s}\right\}\right)\right] + \quad \mathbb{E}_q\left[log\, p\left(\left\{h_{k,n}^{\phi,j,r,s}\right\} \middle| b_k^{\phi,j,r,s}\right)\right] - \mathbb{E}_q\left[log\, q\left(\left\{h_{k,n}^{\phi,j,r,s}\right\}\right)\right]$$

$$+ \quad \mathbb{E}_q\left[log\, p\left(\left\{\theta_k^{\tau,\phi}\right\} \middle| \vartheta_{\tau,\phi}, c\right)\right] - \mathbb{E}_q\left[log\, q\left(\left\{\theta_k^{\tau,\phi}\right\}\right)\right]. \tag{54}$$

The difference between the left and right hand sides of eqn. (54) is the Kullback-Leibler divergence between the true posterior and the variational distribution q. Thus, maximizing this bound with respect to q minimizes the KL divergence between q and our posterior distribution of interest. The likelihood term in eqn. (54) expands to

$$\mathbb{E}_q\left[\log p\left(\left|\Sigma_{r,s,f,n}^{(x)}\right| \left|\left\{w_{f,k}^{\tau,j,r,s}\right\}, \left\{h_{k,n}^{\phi,j,r,s}\right\}, \left\{\theta_k^{\tau,\phi}\right\}\right.\right)\right]$$

$$= \sum_{f,n} \mathbb{E}_q\left[\frac{-\left|\Sigma_{r,s,f,n}^{(x)}\right|}{\sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}}\right] - \mathbb{E}_q\left[\log \sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}\right]. \tag{55a}$$

By using the Jensen's inequality, it can be shown that the above likelihood term is bounded below as:

$$\mathbb{E}_q\left[\log p\left(\left|\Sigma_{r,s,f,n}^{(x)}\right| \left|\left\{w_{f,k}^{\tau,j,r,s}\right\}, \left\{h_{k,n}^{\phi,j,r,s}\right\}, \left\{\theta_k^{\tau,\phi}\right\}\right.\right)\right]$$

$$\geq \sum_{f,n} \sum_{j,k,\tau,\phi} -\left|\Sigma_{r,s,f,n}^{(x)}\right| \left(\varphi_{f,n,k}^{\tau,\phi,r,s}\right)^2 \mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}}\right] - log\left(\omega_{f,n}^{r,s}\right) + 1 - \frac{1}{\omega_{f,n}^{r,s}} \mathbb{E}_q\left[\theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}\right]. \tag{55b}$$

Using Lagrange multipliers, maximizing the lower bound eqn. (54) with eqn. (55) leads to the following optimal $\varphi_{f,n,k}^{\tau,\phi,r,s}$ and $\omega_{f,n}^{r,s}$:

$$\varphi_{f,n,k}^{\tau,\phi,r,s} \propto \mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi} w_{f,k}^{\tau,r,s} h_{k,n}^{\phi,r,s}}\right]^{-1}. \tag{56}$$

and

$$\omega_{f,n}^{r,s} = \sum\nolimits_{j,k,\tau,\phi} \mathbb{E}_q \left[\theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}\right]. \tag{57}$$

The variational distribution parameters can be optimized by differentiating the likelihood function in eqn. (54) and eqn. (55) with respect to its parameters. This yields the following updates:

$$\zeta_{w,f,k}^{\tau,r,s} = a_k^{\tau,j,r,s}\,. \tag{58a}$$

$$\psi_{w,f,k}^{\tau,r,s} = a_k^{\tau,j,r,s} + \sum_{n,\phi} \frac{\mathbb{E}_q \left[\theta_k^{\tau,\phi} h_{k,n-\tau}^{\phi,j,r,s}\right]}{\omega_{f,n}^{r,s}}. \tag{58b}$$

$$\beta_{w,f,k}^{\tau,r,s} = \sum_{n,\phi} \left|\Sigma_{r,s,f,n}^{(x)}\right| \varphi_{f,n,k}^{\tau,\phi,r,s^2} \mathbb{E}_q \left[\frac{1}{\theta_k^{\tau,\phi} h_{k,n-\tau}^{\phi,j,r,s}}\right]. \tag{58c}$$

$$\zeta_{h,k,n}^{\phi,r,s} = b_k^{\phi,j,r,s}. \tag{59a}$$

$$\psi_{h,k,n}^{\phi,r,s} = b_k^{\phi,j,r,s} + \sum_{f,\tau} \frac{\mathbb{E}_q \left[\theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s}\right]}{\omega_{f,n}^{r,s}}. \tag{59b}$$

$$\beta_{h,k,n}^{\phi,r,s} = \sum_{f,\tau} \left|\Sigma_{r,s,f,n}^{(x)}\right| \varphi_{f,n,k}^{\tau,\phi,r,s^2} \mathbb{E}_q \left[\frac{1}{\theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j,r,s}}\right]. \tag{59c}$$

$$\zeta_{\theta,k}^{\tau,\phi} = \frac{\vartheta_{\tau,\phi}}{L + \phi_{max} + \tau_{max}}. \tag{60a}$$

$$\psi_{\theta,k}^{\tau,\phi} = \vartheta_{\tau,\phi} c + \sum_{f,n} \frac{\mathbb{E}_q \left[w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}\right]}{\omega_{f,n}^{r,s}}. \tag{60b}$$

$$\beta_{\theta,k}^{\tau,\phi} = \sum\nolimits_{f,n} \left|\Sigma_{r,s,f,n}^{(x)}\right| \varphi_{f,n,k}^{\tau,\phi,r,s^2} \mathbb{E}_q \left[\frac{1}{w_{f-\phi,k}^{\tau,j,r,s} h_{k,n-\tau}^{\phi,j,r,s}}\right]. \tag{60c}$$

4.2. Estimating the number of effective components

The number of effective components in the proposed model can be estimated according to the hidden latent variable in eqn. (47) as

$$\mathbb{E}_q[\theta_k] = \int \theta_k \, q(\theta_k) d\theta_k \; = \int \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} \theta_k \, q(\theta_k|\tau,\phi) q(\tau) q(\phi) \, d\theta_k$$

$$= \frac{1}{\tau_{max}\phi_{max}} \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} \mathbb{E}_q\left[\theta_k^{\tau,\phi}\right]. \tag{61a}$$

where

$$\mathbb{E}_q\left[\theta_k^{\tau,\phi}\right] = \int \theta_k \, q(\theta_k|\tau,\phi) d\theta_k = \frac{\sqrt{\beta_{\theta,k}^{\tau,\phi}/\psi_{\theta,k}^{\tau,\phi}}\,\mathcal{K}_{\zeta_{\theta,k}^{\tau,\phi}+1}\left(2\sqrt{\psi_{\theta,k}^{\tau,\phi}\beta_{\theta,k}^{\tau,\phi}}\right)}{\mathcal{K}_{\zeta_{\theta,k}^{\tau,\phi}}\left(2\sqrt{\psi_{\theta,k}^{\tau,\phi}\beta_{\theta,k}^{\tau,\phi}}\right)}. \tag{61b}$$

The above statistical expectations are obtained from the GIG distribution. It is assumed that both $q(\tau)$ and $q(\phi)$ are uniformly distributed. We define the *effective component* as

$$k_* = \arg_k \left\{ \mathbb{E}_q[\theta_k] \Big/ \sum_{k=1}^{K} \mathbb{E}_q[\theta_k] \; \geq \varepsilon \right\}. \tag{62}$$

where ε is a small constant (which we set to 0.1 after conducting 200 experimental trials). We select the optimum model for (τ, ϕ) by treating each $\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi}\right]$ for various values of (τ, ϕ) as a histogram. Thus the optimum model for (τ, ϕ) is given by the average of non-zero components:

$$\hat{\tau}_{max,k_*} = \frac{\sum_{l=0}^{\phi_{max}-1} F_l^{(\tau)}}{\#\left(F_l^{(\tau)} \neq 0, \forall l\right)}. \tag{63a}$$

$$\hat{\phi}_{max,k_*} = \frac{\sum_{l=0}^{\tau_{max}-1} F_l^{(\phi)}}{\#\left(F_l^{(\phi)} \neq 0, \forall l\right)}. \tag{63b}$$

where

$$F_l^{(\tau)} = \#component \; in \; \frac{\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi=l}\right]}{\sum_\tau \mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi=l}\right]} \geq \varepsilon,$$

$$F_l^{(\phi)} = \#component \; in \; \frac{\mathbb{E}_q\left[\theta_{k=k_*}^{\tau=l,\phi}\right]}{\sum_\phi \mathbb{E}_q\left[\theta_{k=k_*}^{\tau=l,\phi}\right]} \geq \varepsilon.$$

The term $F_l^{(\tau)}$ counts the number of τ components in the normalized $\mathbb{E}_q\left[\theta_k^{\tau,\phi=l}\right]$ that exceeds ε,and $\#\left(F_l^{(\tau)} \neq 0, \forall l\right)$ counts the number of entries of $F_l^{(\tau)}$ that is non-zero. The same interpretation is applied to $F_l^{(\phi)}$ and $\#\left(F_l^{(\phi)} \neq 0, \forall l\right)$ for determining the model order ϕ_{max}.

4.3. Initialization of CNTF2D

We initialized the spectral and temporal tensors of the proposed CNTF2D depending on $w_{f,k}^{\tau,j,r,s}$ and $h_{k,n}^{\phi,j,r,s}$ obtained from the Gamma-Exponential process as follows:

$$w_{f,k}^{\tau,j,r,s(initial)} = \mathbb{E}_q\left[w_{f,k}^{\tau,j,r,s}\right]. \tag{64a}$$

$$h_{k,n}^{\phi,j,r,s(initial)} = \mathbb{E}_q\left[h_{k,n}^{\phi,j,r,s}\right]. \tag{64b}$$

for the convolutive parameters and number of components that have been obtained from the Gamma-Exponential process. Equations (64a) and (64b) can be obtained using (61b). Table 1 summarizes the main step of the proposed CNTF2D algorithm.

Table 1: Proposed algorithm GEM-MU CNTF2D

Step 1: Estimate the number of components and convolutive parameters by using the proposed Gamma-Exponential process in eqns. (58)-(60) and compute $\mathbb{E}_q\left[\theta_k^{\tau,\phi}\right]$.

Step 2: Initialize $W = \left\{w_{f,k}^{\tau,j,r,s}\right\}$ and $H = \left\{h_{k,n}^{\phi,j,r,s}\right\}$ with the proposed Gamma-Exponential process spectral and temporal tensors using eqn. (64), $\alpha = \left\{\alpha_{k,f,n}^{j,r,s}\right\}$ with zero, $\Sigma_f^{(b)}$ with random nonnegative diagonal matrix, and λ_p with a positive value.

Step 3: (E-step) Compute $\hat{R}_{j,f,n}^{(c)}$, $\hat{\Sigma}_{j,f,n}^{(c)}$, $\hat{c}_{j,f,n}$, $\hat{R}_f^{(b)}$, $\hat{\Sigma}_f^{(b)}$, and $\hat{b}_{f,n}$ using eqns. (17)-(22).

Step 4: (M-step) Compute $w_{f,k}^{\tau,j,r,s}$, $h_{k,n}^{\phi,j,r,s}$, $\alpha_{k,f,n}^{j,r,s}$, and λ_p using eqns. (29), (30),(31), and (39).

Step 5: Normalize $w_{f,k}^{\tau,j,r,s} = w_{f,k}^{\tau,j,r,s} / \sqrt{\sum_{f,k,\tau} \left(w_{f,k}^{\tau,j,r,s} \right)^2}$.

Step 6: Repeat Steps 3 to 5 until convergence is achieved i.e., rate of cost change is below a prescribed threshold, ψ (e.g., $\psi = -20dB$).

Step 7: Perform inverse STFT with dual synthetic window to estimate $c_{i,j}(t)$.

---

5. RESULTS AND DISCUSSIONS

5.1. DATASET

The following two datasets will be used in the experiments.

5.1.1. *Dataset 1:* This dataset is identical to the one used in the full-rank NMF of Arberet et al. algorithm [8]. This dataset consist of four groups depending on the distance between their microphones and the reverberation time (RT). These are the 5 cm distance with 130 ms reverberation time group, 5 cm and 250 ms group, 1 m and 130 ms group, and 1 m 250 ms group. Each group consists of ten stereo mixtures, and each mixture has a length of 10 seconds, sampled at 16 kHz, and generated from three convolutive sources.

5.1.2. *Dataset 2:*This is an under-determined speech and music mixtures development dataset of SiSEC 2018 [32]. This dataset consist of two groups. The first group is the live recording music group, which consists of dev1 and dev2 datasets, where each dataset has the with drum (wdrum) group; which consists of vocal and music instrument with drum, and the without drum (nodrum) group; which consists of vocal and music instruments without drum. The sources of this group are mixed in stereo mixture that has 1 m or 5 cm space between its microphones, and 250 ms reverberation time. The second group of this dataset is a simulated recording speech group, which consists of dev3 dataset, this dataset contains four females (female4) and four males (males4) that mixed in stereo mixture, with 5 cm or 50 cm distance between its microphones, and has a

reverberation time of 130 ms or 380 ms. Dev3 has three channels (left, right, and mono) and we reduce it to two channels (left and right). Additionally, each mixture has duration of 10 s and sampled at 16 kHz.

5.2. Evaluation

The performance of the proposed algorithm will be measured by using the signal-to-distortion ratio (SDR) [33] which measures an overall sound quality of the source separation, where it combines the signal-to-interference ratio (SIR), source image-to-spatial distortion ratio (ISR), and the signal-to-artifact ratio (SAR), into one measurement.

5.3. Effects of Variable Sparsity versus Uniform Sparsity

In this subsection, we will show the effects of the sparsity on the separation performance, by considering a fixed uniform sparsity; $\lambda_{k,n}^{\phi,j,r,s} = \lambda = c$, all over the elements of H, and the variable sparsity $\lambda_{k,n}^{\phi,j,r,s}$ for each element of H. The fixed uniform sparsity is commonly used throughout the literature of matrix factorization. Each experiment will be run for different values of sparsity for the three sources that convolutively mixed in the stereo mixture that has 1 m space between its microphones, 250ms reverberation time, and with 16 kHz sampling frequency. The following parameters are set for the proposed algorithm: $K_j = 4$, $\tau_{max} = 3$, and $\phi_{max} = 3$. In order to focus on the effects of sparsity, oracle initialization has been used. Fig. 3 shows the average SDR performance with respect to different values of sparsity. The variable sparsity has resulted in the highest SDR performance. This is attributed to the fact that each element of *H* has a specific sparsity value instead of constant value for the entire set of *H* as in the case of uniform sparsity. This is especially more pronounced in audio signals in which case the spectrogram has a large dynamic range. It is seen that for variable sparsity the average SDR is 3.2 dB higher than the best uniform sparsity (the value of constant λ that results in the highest SDR) $\lambda = 5$. Additionally, as the sparsity value increases (leading to over-sparseness) the SDR begins to decrease since many elements in *H* become very small and tend to zero. This results in switching off several parts of the spectrum in the estimated sources, as shown in Fig. 4. In particular, it shows the spectrogram of one of the estimated sources for the case of variable sparsity, over-sparse, and under-sparse.

It is visually apparent that the over-sparse and under-sparse have not fully recovered the original source. Many parts of the spectrum have been removed from the estimated source due to over-sparseness of H while many unwanted spectrum have been added to' the estimated source with under-sparseness. On the other hand, the variable sparsity has resulted in almost full recovery of the original source, as it has been optimally tuned by the degree of sparseness over all the elements of H.

5.4. Separation Results

5.4.1. Estimation of number of components and convolutive parameters

The proposed Gamma-Exponential CNTF2D process has been applied to the mixtures of Dataset 1 and Dataset 2, and the estimated values are tabulated in Table 2 for Dataset 1, and in Tables 4 and 5 for Dataset 2. It can be seen from these tables that we have different parameters ($\tau, \phi,$ and K) for each mixture as each mixture has a different temporal and pitches characteristics. In the following, we detail an example from Dataset 1 on how the model order is selected. Firstly, we set the bound of the proposed Gamma-Exponential process as follows: $\tau = \{0, 1, 2, \ldots, 10\}$, $\phi = \{0,1,2,\ldots,10\}$, and $K = 20$. Secondly, we run the proposed model order estimation step (eqns. (56)-(63)) and the results of the Gamma-Exponential process are shown in Fig. 5. Thirdly, we estimate the effective parameters ($\tau, \phi,$ and K) in Fig. 5. Fig. 5(a) shows the values of $\mathbb{E}_q[\theta_k]$ for $k = 1, \ldots, 20$ which are predominantly zero except for $k = 3, 5, 8, 10, 11, 12, 13, 16, 17, 18$ and 19. Let $K_* = \#\, k_*$ be the number of effective components; from Fig. 5(a) this corresponds to $K_* = 11$. Since there are $J = 3$ sources, then $K_j = K_*/J \approx 4$ for $j = 1, 2, 3$. In addition, for each k_* effective component, we have determined distribution for (τ, ϕ) which is given by $\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi}\right]$. These are shown in Fig. 5(b)-(l): $\left(\hat{\tau}_{max,3} = 4, \hat{\phi}_{max,3} = 2\right)$, $\left(\hat{\tau}_{max,5} = 1, \hat{\phi}_{max,5} = 6\right)$, $\left(\hat{\tau}_{max,8} = 2\ \hat{\phi}_{max,8} = 4\right)$, $\left(\hat{\tau}_{max,10} = 2\ \hat{\phi}_{max,10} = 3\right)$, $\left(\hat{\tau}_{max,11} = 3\ \hat{\phi}_{max,11} = 3\right)$, $\left(\hat{\tau}_{max,12} = 2\ \hat{\phi}_{max,12} = 4\right)$, $\left(\hat{\tau}_{max,13} = 2\ \hat{\phi}_{max,13} = 4\right)$, $\left(\hat{\tau}_{max,16} = 3\ \hat{\phi}_{max,16} = 2\right)$, $\left(\hat{\tau}_{max,17} = 2\ \hat{\phi}_{max,17} = 3\right)$, $\left(\hat{\tau}_{max,18} = 1\ \hat{\phi}_{max,18} = 7\right)$ *and* $\left(\hat{\tau}_{max,19} = 2\ \hat{\phi}_{max,19} = 2\right)$, respectively, and its averages are $\hat{\tau}_{max} = \frac{1}{K_*}\sum_{k_*} \hat{\tau}_{max,k_*} = 2$, and $\hat{\phi}_{max} = \frac{1}{K_*}\sum_{k_*} \hat{\phi}_{max,k_*} = 3$.

5.4.2. Results of Dataset 1

In this dataset, the STFT window length is set to 1024 with 50% overlaps, and 50 iterations are used for testing the competing algorithms. For comparison purposes, we used the same initialization as that used in Arberet et al.. Furthermore, as the oracle initialization is used there will be no further need to include the phase, so we set α to zero. To show the convergence of the proposed algorithm, the average cost functions in eqn. (15) of the ten mixtures with different conditions (low and high reverberations time, and short and long distance between the microphones) are shown in Fig. 6. It is noted that the speed of convergence (as measured by the gradient of the cost function) is fastest for the short microphone distance with low reverberation. As the microphone distance becomes larger and the level of reverberation increases, the speed tends to slow down. Nonetheless, all curves converge to the steady state in less than 50 iterations. Furthermore, the SDRs of the full-rank NMF and the proposed algorithm are tabulated in Table 3. The table indicates that the proposed algorithm has better performance than the full-rank NMF since it has more powerful representation (using the CNTF2D), as well as the variable sparsity over all the elements of H. We summarize the results for all the conditions as follows: An average achievement of 1.2 dB more for the low reverberation group, and an average of 0.9 dB more on average for the high reverberations group. It shows that high SDR performance has been achieved for the 130ms reverberation for both 100cm and 5cm microphone separation. This case corresponds to the low reverberation environment. For the case of high reverberation, the proposed algorithm performs better with shorter microphone distance. As the distance between the microphones decreases, the signal at each microphone becomes more correlated with each other and therefore the channel covariance matrix $\Sigma_{j,f}^{(a)}$ in eqn. (4a) tends to have specific structure and hence reinforces the requirement of full-rank condition. On the other hand, as the separation between the microphones increases, the signal at each microphone becomes less correlated with each other. The effect is that each channel behaves independently and the channel covariance matrix $\Sigma_{j,f}^{(a)}$ can be modelled by rank-1 structure. Thus as the separation between microphone becomes progressively small, this induces a complex structure to the channel covariance which will benefit from the

full-rank estimation procedure in the proposed algorithm. There is a clear indication that the proposed algorithm has outperformed the NMF for both the low and high reverberation time. The spectrogram of one of the original sources, and its estimate by using the full-rank NMF and the variable sparsity CNTF2D are shown in Fig. 7(a), (b), and (c), respectively. These figures clearly show that the variable sparsity CNTF2D has successfully detected the pitch change of the source (as shown in the high frequency of its spectrogram), due to its two-dimensional deconvolution while the full-rank NMF failed to detect these changes. Furthermore, one component of the W and H matrices and its corresponding reconstructed spectrogram is shown in Fig. 7(d). This clearly indicates that both W and H have modelled the sources quite accurately. It is seen that W has successfully modelled the frequencies of the source especially in the high frequency region and H has shown a correct distribution in the time domain.

5.4.3. Result of Dataset 2

In this section, we compare our algorithm with Adiloglu's work in [34] from the SiSEC'16 evaluation campaign for the tasks of under-determined speech and music mixtures which uses fully Bayesian source separation algorithm based on variational inference method [35], with the multi-level NMF model [36] as a source variance, and the time difference of arrival (TDOA) as an initialization method [37]. Also we compare our algorithm with the standard NTF2D optimized using Euclidean distance [21]. The STFT window length is set to 2048 with 50% overlaps. Furthermore, for fair comparison and to show the significance of the convolutive parameters, we set the convolutive parameters of the proposed algorithm to zero. In other words, we compare with the full rank complex NTF instead of the NTF2D. We term this algorithm as the GEM-MU variable sparsity complex NTF. The average cost functions are shown in Fig. 8. The figure indicates that all the cost functions converged to a low value within 10 iterations while Adiloglu's algorithm requires about 250 iterations. Furthermore, Table 4 shows the SDRs of the proposed algorithm for the music group and on average it yields higher value than Adiloglu's algorithm. For clarity of comparison, the results are summarized as follows: An improvement of 2.5 dB is achieved for the 5 cm and 100 cm distance with 250 ms reverberation

time datasets. Table 5 shows the results for the speech group and on average an improvement of 2.5 dB has been achieved for the 5 cm, 380 ms datasets, 1.9 dB for the 50 cm, 380 ms datasets, 0.3 dB for the 5 cm, 130 ms datasets, and 0.1 dB for the 50 cm, 130 ms datasets. For the NTF2D, the SDRs of the proposed algorithm are better for all the cases. Finally, Fig. 9 shows the spectrogram of the estimated sources. It has indicated that the proposed algorithm has successfully estimated the sources with a reasonable degree of accuracy. In particular, it is evident that all the low and high frequency components as well as the time-frequency patterns have been preserved in the estimated sources.

5.4.4. Robustness to Noise and Computational Complexity

Two additional assessments of the proposed method have been undertaken to clarify on the computational complexity and robustness against noise. Using the SiSEC 2016: Dev. 2 dataset running on a PC with dual core processor @ 2.4 GHz (i7 Intel processor) 8 GB RAM and 320 GB HDD, the computational time taken by each algorithm has been tabulated in Table 6. It is shown that the time taken to run one iteration is the highest for the proposed algorithm. However, the proposed algorithm has fastest convergence to the steady state solution requiring on average about 41 iterations. Comparing in terms of the total computational time, the proposed algorithm is computationally more demanding than Adiloglu's algorithm and NTF2D by 19.9% and 12.4%, respectively. This is due to the estimation of the sparsity parameter which is computationally most demanding. We have also performed a test to examine the robustness of the algorithms in separating the mixture under different level of signal-to-noise ratio (SNR). Fig. 10 shows the obtained result using the SiSEC 2016: Dev. 2 dataset. It is shown that the proposed algorithm has consistently outperformed the Adiloglu's algorithm and NTF2D. Note that the measured SDR here correspond to the average of SDR of all separated source images. In addition, the proposed algorithm has been able to maintain a graceful depreciation of the SDR as the SNR reduces. This is attributed to the ability of the algorithm in modelling the noise and realizing the conditional estimate of the source images via the Wiener filter. Overall, it can be deduced that although the proposed

algorithm has higher computational load, its separation performance as measured by the SDR shows it is robust against noise.

6. CONCLUSIONS

In this paper, a novel method that combines the complex NTF2D model with variable sparsity has been proposed for multichannel source separation. The variable sparse parameters are derived from the Gibbs distribution, which has provided a tractable and stable approach to adapt each sparse parameter for every temporal code in the CNTF2D. The GEM-MU algorithm has been used as a platform to enable the joint estimation of the sources and parameters as well as preserving the non-negativity constraints of the proposed algorithm. It outperforms the full-rank NMF and NTF algorithms, and a recent algorithm based on variational inference multi-level NMF model with TDOA initialization. The proposed algorithm is fast and requires less than 10 iterations to converge to the steady state. Finally, the parameters that affect the separation such as initialization, convolutive parameters, and number of components have been controlled by using the proposed Gamma-Exponential process to minimize the randomization in the separation algorithm.

APPENDIX

A1. Derivations of variational parameter u_l

The distribution $Q_L(\underline{h}_L)$ in (40) will be approximated by considering the Taylor expansion about the updated h^*:

$$Q_L(\underline{h}_L \geq 0) \propto exp\left\{-\sum_{l\in L}\left(\left(\frac{\partial F(h_l)}{\partial h_l}\right)\Big|_{h^*}\right)h_l - \frac{1}{2}\sum_{l\in L}\left(\left(\frac{\partial^2 F(h_l)}{\partial h_l^2}\right)\Big|_{h^*}\right)h_l^2\right\}$$

$$\propto exp\left\{\sum_{r,s,jkf\phi l}\left(\hat{R}^{(c)}_{r,s,j,f,l}\Sigma^{(c)^{-2}}_{r,s,j,f,l}\,e^{-\sqrt{-1}\,\alpha^{j,r,s}_{f,l}}w^{\tau,j,r,s}_{f-\phi,k}-\Sigma^{(c)^{-1}}_{r,s,j,f,l}\,e^{-\sqrt{-1}\,\alpha^{j,r,s}_{f,l}}w^{\tau,j,r,s}_{f-\phi,k}\;-\;\lambda_l\right)h_l\right.$$

$$\left.+\frac{1}{2}\sum_{r,s,jkf\phi l}\left(-2\left(w^{\tau,j,r,s}_{f-\phi,k}\right)^2 e^{-2\sqrt{-1}\,\alpha^{j,r,s}_{f,l}}\hat{R}^{(c)}_{r,s,j,f,l}\Sigma^{(c)^{-3}}_{r,s,j,f,l}+\left(w^{j}_{f-\phi,k}\right)^2 e^{-2\sqrt{-1}\alpha^{j,r,s}_{f,l}}\Sigma^{(c)^{-2}}_{r,s,j,f,l}\right)h_l^2\;\right\}. \quad (65)$$

The variational approximation of $Q_L(\underline{h}_L)$ will be considered by the exponential distribution:

$$\hat{Q}_p(\underline{h}_L \geq 0) = \prod_{l\in L}\frac{1}{u_l}exp\left(-\frac{h_l}{u_l}\right). \quad (66)$$

The parameter u_l is obtained by minimizing the Kullback- Leibler divergence between Q_L and $\hat{Q}_L$

$$u_l = \arg\min_{u_l}\int \hat{Q}_L(\underline{h}_L)\log\frac{\hat{Q}_p(\underline{h}_L)}{Q_p(\underline{h}_L)}\,d\underline{h}_L. \quad (67)$$

where

$$\int \hat{Q}_L(\underline{h}_L)[ln\,\hat{Q}_L(\underline{h}_L)]d\underline{h}_L = \sum_{l\in L}\int_0^\infty \frac{1}{u_l}exp\left(-\frac{h_l}{u_l}\right)\left(-ln\,u_l-\frac{h_l}{u_l}\right)dh_l$$

$$= -\sum_{l\in L} ln\,u_l + 1. \quad (68)$$

and

$$\int \hat{Q}_L(\underline{h}_L)ln\,Q_L(\underline{h}_L)\;d\underline{h}_L$$

$$= \sum_{r,s,j,k,f,\phi,l}\left(\hat{R}^{(c)}_{r,s,j,f,l}\Sigma^{(c)^{-2}}_{r,s,j,f,l}\,e^{-\sqrt{-1}\,\alpha^{j,r,s}_{f,l}}w^{\tau,j,r,s}_{f-\phi,k}-\Sigma^{(c)^{-1}}_{r,s,j,f,l}\,e^{-\sqrt{-1}\,\alpha^{j,r,s}_{f,l}}w^{\tau,j,r,s}_{f-\phi,k}\;-\;\lambda_l\right)u_l$$

$$+\frac{1}{2}\sum_{r,s,j,k,f,\phi,l}\left(-2\left(w^{\tau,j,r,s}_{f-\phi,k}\right)^2 e^{-2\sqrt{-1}\,\alpha^{j,r,s}_{f,l}}\hat{R}^{(c)}_{r,s,j,f,l}\Sigma^{(c)^{-3}}_{r,s,j,f,l}+\left(w^{j}_{f-\phi,k}\right)^2 e^{-2\sqrt{-1}\alpha^{j,r,s}_{f,l}}\Sigma^{(c)^{-2}}_{r,s,j,f,l}\right)u_l u_l\,. \quad (69)$$

Thus,

$$u_l = \arg\min_{u_l} \left(-\sum_{l\in L} ln u_l + 1 + \sum_{r,s,j,k,f,\phi,l} \left(\hat{R}^{(c)}_{r,s,j,f,l} \Sigma^{(c)^{-2}}_{r,s,j,f,l} e^{-\sqrt{-1}\,\alpha^{j,r,s}_{f,l}} w^{\tau,j,r,s}_{f-\phi,k} - \Sigma^{(c)^{-1}}_{r,s,j,f,l} e^{-\sqrt{-1}\,\alpha^{j,r,s}_{f,l}} w^{\tau,j,r,s}_{f-\phi,k} - \lambda_l \right) u_l \right.$$

$$\left. + \frac{1}{2} \sum_{r,s,j,k,f,\phi,l} \left(-2\left(w^{\tau,j,r,s}_{f-\phi,k}\right)^2 e^{-2\sqrt{-1}\,\alpha^{j,r,s}_{f,l}} \hat{R}^{(c)}_{r,s,j,f,l} \Sigma^{(c)^{-3}}_{r,s,j,f,l} + \left(w^{\tau,j,r,s}_{f-\phi,k}\right)^2 e^{-2\sqrt{-1}\alpha^{j,r,s}_{f,l}} \Sigma^{(c)^{-2}}_{r,s,j,f,l} \right) u_l . u_l \right). \quad (70)$$

Let $\tilde{\Theta}, \Theta_l$, and b_l be defined according to eqns. (41)-(42), then we have

$$u_l = \arg\min_{u_l} \left(\underline{b}^H_L \underline{u} + \frac{1}{2} \underline{u}^H \tilde{\Theta} \underline{u} - \sum_{l\in L} \ln u_l \right). \quad (71)$$

By using the nonnegative quadratic programming (NQP) [38], we have

$$G(\underline{u}, \tilde{\underline{u}}) = \underline{b}^H_L \underline{u} + \frac{1}{2} \sum_{l\in L} \frac{\left(\tilde{\Theta}\tilde{\underline{u}}\right)_l}{\tilde{u}_l} u_l^2 - \sum_{l\in L} \ln u_l . \quad (72)$$

Taking the derivative of $G(u, \tilde{u})$ in eqn. (72) with respect to *u* and setting it to zero yields

$$\frac{\left(\tilde{\Theta}\tilde{\underline{u}}\right)_l}{\tilde{u}_l} u_l^2 + \underline{b}^H_L u_l - 1 = 0. \quad (73)$$

which is solved as in eqn. (40).

REFERENCES

[1] W.L. Woo, B. Gao, A. Bouridane, B.W-K Ling, C.S. Chin, "Unsupervised Learning with Monaural Source Separation using Maximum-Minimum Algorithm and Time-Frequency Deconvolution," *Sensors*, vol. 18, no. 5, 2018.

[2] Ahmed Al-Tmeme, W.L. Woo, S.S. Dlay and B. Gao, “Underdetermined Convolutive Source Separation using GEM-MU with Variational Approximated Optimum Model Order NMF2D,” *IEEE Trans. Audio Speech and Language Processing,* vol. 25, no. 1, pp. 35-49, 2017.

[3] A. Ozerov, and C. Fevotte, “Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation,” *IEEE Trans. Audio Speech and Language Processing,* vol. 18, no. 3, pp. 550-563, Mar, 2010.

[4] N.Q.K. Duong, E. Vincent, and R. Gribonval, “Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model,” *IEEE Trans. Audio Speech and Language Processing*, vol. 18, no. 7, pp. 1830-1840, Sep, 2010.

[5] N.Q.K. Duong, E. Vincent, and R. Gribonval, “Under-Determined Reverberant Audio Source Separation Using Local Observed Covariance and Auditory-Motivated Time-Frequency Representation,” in *9th International Conference on Latent Variable Analysis and Signal Separation* (LVA/ICA’10), 2010, pp. 73-80.

[6] M. Taoufiki, A. Adib, and D. Aboutajdine, “A new behavior of higher order blind source separation methods for convolutive mixture,” *Digital Signal Processing*, vol. 20, no. 1, pp. 269-275, 2010.

[7] Ahmed Al-Theme, W.L. Woo, S.S. Dlay, and B. Gao, “Underdetermined reverberant acoustic source separation using weighted full-rank nonnegative tensor models,” *J. Acoust. Soc. Am*, 138, 3411, 2015.

[8] Ahmed Al-Theme, W.L. Woo, S.S. Dlay, and B. Gao, “Single Channel Informed Signal Separation using Artificial-Stereophonic Mixtures and Exemplar-Guided Matrix Factor Deconvolution,” *International Journal Adaptive Control and Signal Processing*, 2018, https://doi.org/10.1002/acs.2912.

[9] D. Peng and Y. Xiang, "Underdetermined blind separation of non-sparse sources using spatial time-frequency distributions," *Digital Signal Processing*, vol. 20, no. 2, pp. 581-596, 2010.

[10] N. Tengtrairat, W.L. Woo, S.S. Dlay, and B. Gao, “Online Noisy Single-Channel Blind Separation by Spectrum Amplitude Estimator and Masking,” *IEEE Trans. Signal Processing*, vol. 64, no. 7, pp. 1881-1895, 2016.

[11] N. Tengtrairat, Bin Gao, W.L. Woo and S.S. Dlay, “Single-Channel Blind Separation using Pseudo-Stereo Mixture and Complex 2-D Histogram,” *IEEE Trans. on Neural Networks and Learning Systems*, vol. 24, no. 11, pp. 1722-1735, 2013.

[12] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, “Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation,” in *Information Sciences Signal Processing and their Applications (ISSPA)*, 2010 10th International Conference on, 2010, pp. 1-4.

[13] B. J. King, and L. Atlas, “Single-Channel Source Separation Using Complex Matrix Factorization,” *IEEE Trans. Audio Speech and Language Processing*, vol. 19, no. 8, pp. 2591-2597, Nov, 2011.

[14] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *7th International Symposium on Computer Music Modeling and Retrieval (CMMR 2010)*, Málaga, Spain, 2010, www.cmmr2010.etsit.uma.es.

[15] T. Barker and T. Virtanen, "Non-negative Tensor Factorisation of Modulation Spectrograms for Monaural Sound Source Separation," in *Proc. of 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), 25-29 August, Lyon, France*, ed: International Speech Communication Association, 2013, pp. 827-831.

[16] T. Barker and T. Virtanen, "Blind Separation of Audio Mixtures Through Nonnegative Tensor Factorization of Modulation Spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 2377-2389, 2016.

[17] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. of 22nd international conference on Machine learning*, Bonn, Germany, 2005.

[18] B. Gao, W.L. Woo and S.S. Dlay, "Variational Regularized Two-Dimensional Nonnegative Matrix Factorization," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no.5, pp. 703-716, 2012.

[19] C. Fevotte, N. Bertin, and J. L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation,* vol. 21, no. 3, pp. 793-830, Mar, 2009.

[20] B. Gao, W.L. Woo, and L.C. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," *J. Acoust. Soc. Am.*, vol. 135, pp. 1171-1185, 2014.

[21] M. Morup, and M. N. Schmid, *Sparse Non-Negative Matrix Factor 2-D Deconvolution*, Tech. Rep Technical University of Denmark, Copenhagen, Denmark, 2006.

[22] B. Gao, W. L. Woo, and S. S. Dlay, "Nonnegative matrix factorization for single channel source separation " *IEEE Journal of Selected Topics in Signal Processing,* vol. 5, no. 5, pp. 989-1001, 2011.

[23] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised Single-Channel Separation of Nonstationary Signals Using Gammatone Filterbank and Itakura-Saito Nonnegative Matrix Two-Dimensional Factorizations," *IEEE Trans. Circuits and Systems I-Regular Papers,* vol. 60, no. 3, pp. 662-675, Mar, 2013.

[24] M. N. Schmidt, and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *6th Intl. Conf. on Independent Component Analysis and Signal Separation (ICA '06)*, Charleston, USA, 2006, pp. 700–707.

[25] P. Parathai, W.L. Woo, S.S. Dlay and B. Gao, "Single-Channel Blind Separation using L1-Sparse Complex Nonnegative Matrix Factorization for Acoustic Signals," *J. Acoust. Soc. Am*, 137, EL124, 2015.

[26] F.D. Neeser, and J.L. Massey, "Proper Complex Random-Processes with Applications to Information-Theory," *IEEE Trans. Information Theory,* vol. 39, no. 4, pp. 1293-1302, Jul, 1993.

[27] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Nov. 2004.

[28] A. Abdullah, J. Moeller, and S. Venkatasubramanian, “Approximate Bregman near Neighbors in Sublinear Time: Beyond the Triangle Inequality,” *International Journal of Computational Geometry & Applications,* vol. 23, no. 4-5, pp. 253-301, Aug-Oct, 2013.

[29] Qi Wang, W.L. Woo, and S.S. Dlay, "Informed Single Channel Speech Separation Using HMM-GMM User-Generated Exemplar Source," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2087-2100, 2014.

[30] M. Hoffman, D. Blei, and P. Cook, “Bayesian nonparametric matrix factorization for recorded music,” in *International Conference on Machine Learning (ICML)*, 2010, pp. 439-446.

[31] P. Embrechts, “A Property of the Generalized Inverse Gaussian Distribution with Some Applications,” *Journal of Applied Probability,* vol. 20, no. 3, pp. 537-544, 1983.

[32] "Signal Separation Evaluation Campaign (SiSEC 2018)," 2018; https://sisec.wiki.irisa.fr/ (date last viewed 02/05/18).

[33] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results,” in *Independent Component Analysis and Signal Separation: 7th International Conference, ICA 2007,* 2007, pp. 552-559.

[34] K. Adiloglu, H. Kayser, and L. Wang, "A variational inference based source separation approach for the separation of sources in underdetermined recording," (2013); http://www.onn.nii.ac.jp/sisec16/evaluation_result/UND/submission/ob/Algorithm.pdf (date last viewed 01/06/17).

[35] K. Adiloglu, and E. Vincent, *“Variational Bayesian interference for source separation and robust feature extraction,* Tech. Rep. RT-0428, Inria, Augest 2012.

[36] A. Ozerov, E. Vincent, and F. Bimbot, “A General Fexible Framework for the Handling of Prior Information in Audio Source Separation,” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20 no. 4, pp. 1118–1133, May, 2012.

[37] C. Knapp, and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320-327, 1976.

[38] B. Gao, W.L. Woo and S.S. Dlay, “Single Channel Blind Source Separation Using EMD-Subband Variable Regularized Sparse Features,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 961-976, Mar, 2011.

Table 2
Number of components and convolutive parameters for mixtures 1 to 10

Mixture	K_j	$\hat{\tau}_{max}$	$\hat{\phi}_{max}$
1	3	3	2
2	4	3	3
3	4	2	3
4	4	2	3
5	4	1	5
6	5	4	3
7	5	5	3
8	4	3	2
9	4	3	3
10	4	5	2

Table 3
Average SDRs of the 10 mixtures with different conditions for the full-rank NMF and the proposed algorithm

Dataset 1 Reverberation Time (ms)	130		250	
Microphone Distance (cm)	5	100	5	100
SDR of Full-Rank NMF	8.7	10	8.6	9.1
SDR of the proposed algorithm	10.1	11	9.6	9.9

Table 4
SDRs of Adiloglu et al. algorithm, NTF2D, and the proposed algorithms for dev1 and dev2.

Dataset 2			SiSEC 2016: Dev. 1				SiSEC 2016: Dev. 2			
			ndrums		wdrums		ndrums		wdrums	
Reverberation Time (ms)			250		250		250		250	
Microphone Distance (cm)			5	100	5	100	5	100	5	100
Adiloglu et al. Algorithm [31]	SDR	s_1	-5.5	-0.6	7.0	2.4	1.8	4.7	3.7	4.8
		s_2	-1.2	-0.0	-0.1	3.0	2.7	2.0	3.7	2.0
		s_3	3.7	0.6	-0.5	-11.1	-11.7	-3.9	3.7	2.7
		Avg	-2.2	0.0	2.1	-1.9	-2.4	0.9	3.7	3.2
Proposed CNTF with variable sparsity subject to constraint that $\tau_{max} = \phi_{max} = 1$ in (8)	K_j		1		4		4		5	
	SDR	s_1	-0.2	1.2	4.7	4.0	6.9	5.2	1.1	2.0
		s_2	0.6	1.0	1.3	1.4	-1.2	0.1	2.3	0.9
		s_3	1.1	2.1	0.0	0.4	0.4	-2.1	3.3	3.5
		Avg	0.5	1.4	2.0	1.9	2.0	1.1	2.2	2.1
	τ_{max}		10		1		1		7	
	ϕ_{max}		10		6		6		1	
NTF2D [22]	K_j		1		1		1		1	
	SDR	s_1	-4.9	0.2	0.9	8.3	2.9	2.6	-3.3	-3.2
		s_2	1.3	1.9	-4.8	-8.7	-9.5	-3.2	-3.0	-0.6
		s_3	-3.5	1.3	-3.0	2.6	-11.4	-6.4	0.2	-2.4
		Avg	-2.4	1.1	-2.3	0.7	-6.0	-2.3	-2.0	-2.1
Proposed CNTF2D with variable sparsity and *optimized* model order	K_j		1		4		4		5	
	SDR	s_1	1.2	3.3	7.9	7.4	9.3	5.7	2.0	2.7
		s_2	1.9	2.2	1.6	1.7	0.6	1.7	3.9	3.0
		s_3	1.8	3.4	-0.8	0.4	0.6	0.2	3.8	4.7
		Avg	1.6	3.0	2.9	3.2	3.5	2.5	3.2	3.5

Table 5
SDRs of Adiloglu et al., NTF2D, and the proposed algorithms of dev 3.

Dataset 2 SiSEC 2016: Dev. 3			male4				female4			
Reverberation Time (ms)			380		130		380		130	
Microphone Distance (cm)			5	50	5	50	5	50	5	50
Adiloglu et al. Algorithm [31]	SDR	s_1	0.4	-1.7	-2.6	-2.1	0.2	-0.2	-0.0	-1.2
		s_2	-2.6	-0.9	-0.2	2.6	0.2	-1.0	-0.9	0.6
		s_3	-2.1	0.8	1.5	0.8	-3.1	-2.4	0.4	1.4
		s_4	0.0	-0.4	5.2	3.9	-2.8	0.1	4.1	4.4
		Avg	-1.1	-0.6	1.0	1.3	-1.4	-0.9	0.9	1.3
Proposed CNTF with variable sparsity subject to constraint that $\tau_{max} = \phi_{max} = 1$ in (8)	K_j		2				4			
	SDR	s_1	-0.0	0.2	0.8	-0.3	0.3	0.7	0.2	0.5
		s_2	-0.8	-3.6	-1.0	1.0	0.9	0.5	1.6	-0.7
		s_3	0.2	0.3	0.9	-0.5	-0.6	-0.2	0.5	0.9
		s_4	0.8	0.0	0.9	-0.4	0.2	0.2	1.3	-1.2
		Avg	0.0	-0.8	0.4	-0.1	0.2	0.3	0.9	-0.1
	τ_{max}		10				2			
	ϕ_{max}		10				8			
NTF2D [22]	K_j		1				1			
	SDR	s_1	-5.1	-6.6	-3.8	-9.6	-1.6	-5.7	-9.5	-6.8
		s_2	-6.7	-5.7	1.2	2.1	-4.6	-7.7	-2.7	-3.9
		s_3	-7.8	-1.0	-3.9	-3.9	-3.4	-3.4	-5.1	0.6
		s_4	-7.0	-8.4	-5.9	-6.9	-3.6	-2.4	-5.3	-6.6
		Avg	-6.7	-5.4	-3.1	-4.6	-3.3	-4.8	-5.6	-4.2
Proposed CNTF2D with variable sparsity and *optimized* model order	K_j		2				4			
	SDR	s_1	1.2	0.5	1.1	0.5	1.9	1.0	1.4	1.0
		s_2	1.2	1.0	1.3	2.6	0.8	1.1	1.6	1.0
		s_3	1.3	2.2	1.2	0.9	1.3	0.7	0.8	2.9
		s_4	1.4	0.8	1.2	1.2	0.7	1.9	1.3	0.9
		Avg	1.3	1.1	1.2	1.3	1.2	1.2	1.3	1.5

Table 6
Computational time

Algorithm	**Time (s) per iteration**	**Average number of iteration to reach steady-state solution**	**Total time (s)**
Adiloglu	2.35	150	352.5
NTF2D	3.42	110	376.2
Proposed algorithm	10.31	41	422.7

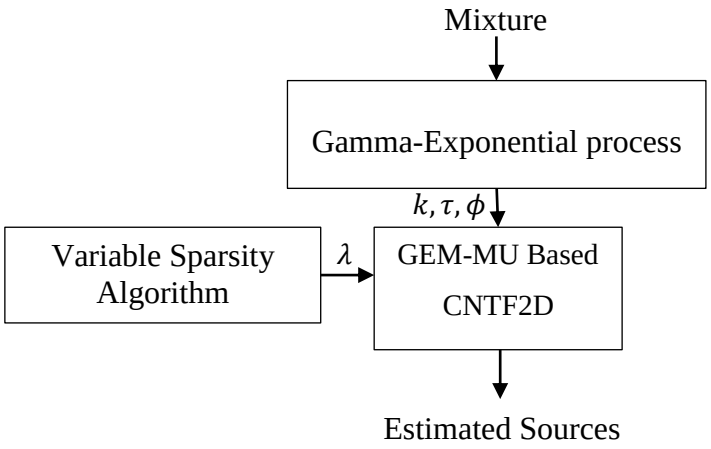

Fig. 1: High level presentation of the proposed system.

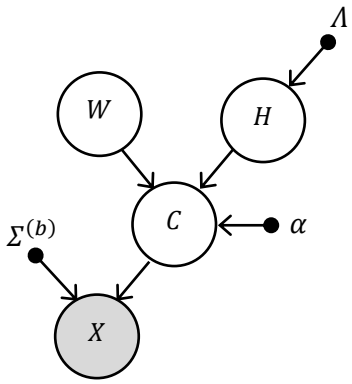

Fig. 2: Graphical model of the proposed CNTF2D.

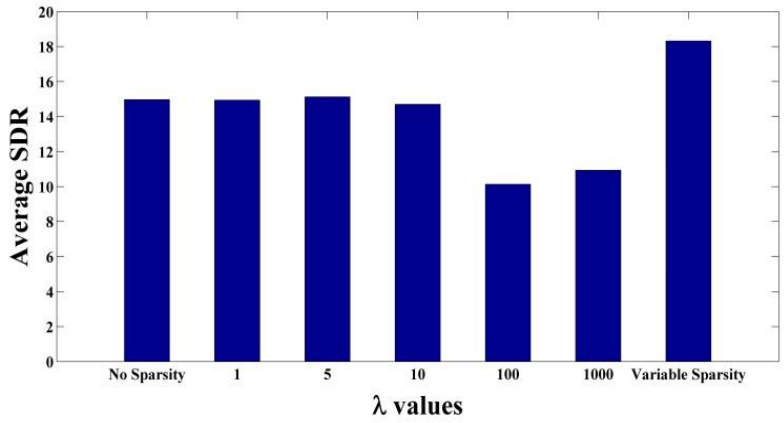

Fig. 3: Average SDR with respect to different sparsity

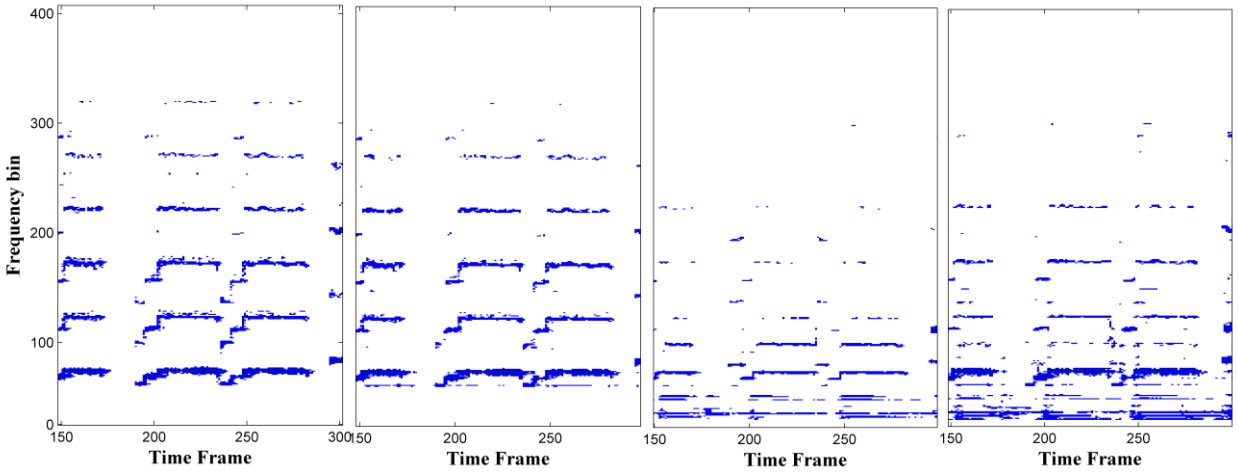

Fig. 4: The effects of sparsity on the estimated source. (a) Original source image. (b) Estimated source with variable sparsity. (c) Estimated source with uniform over-sparse. (d) Estimated source with uniform under-sparse.

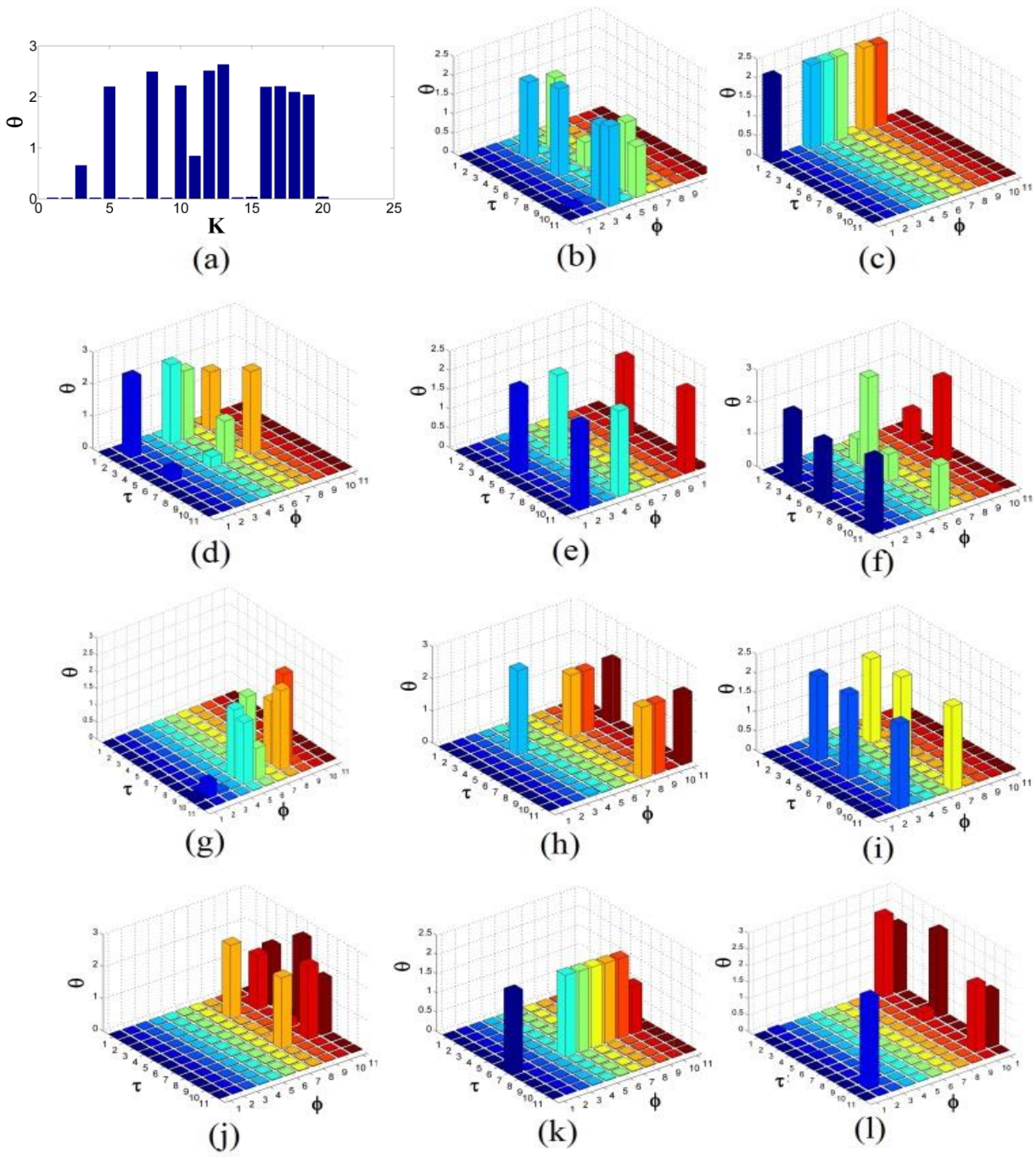

Fig. 5: Estimation of the convolutive parameters and number of components by using the proposed Gamma-Exponential process algorithm. (a) Number of components, (b)–(l) Convolutive parameters corresponding to each component in (a).

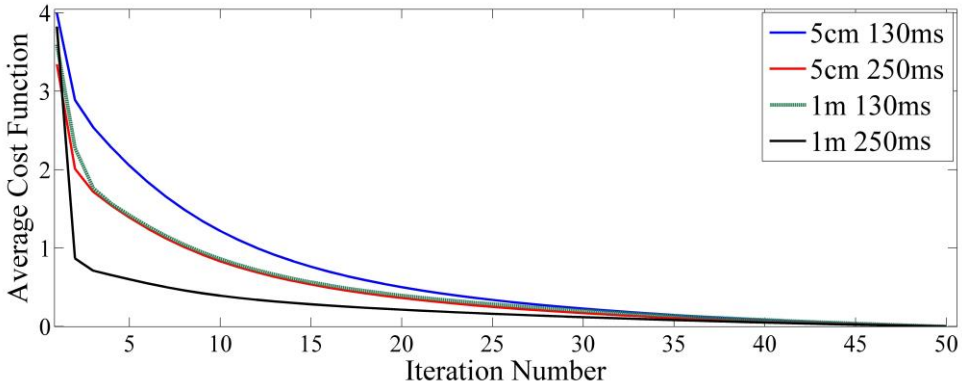

Fig. 6: Average cost function for different conditions.

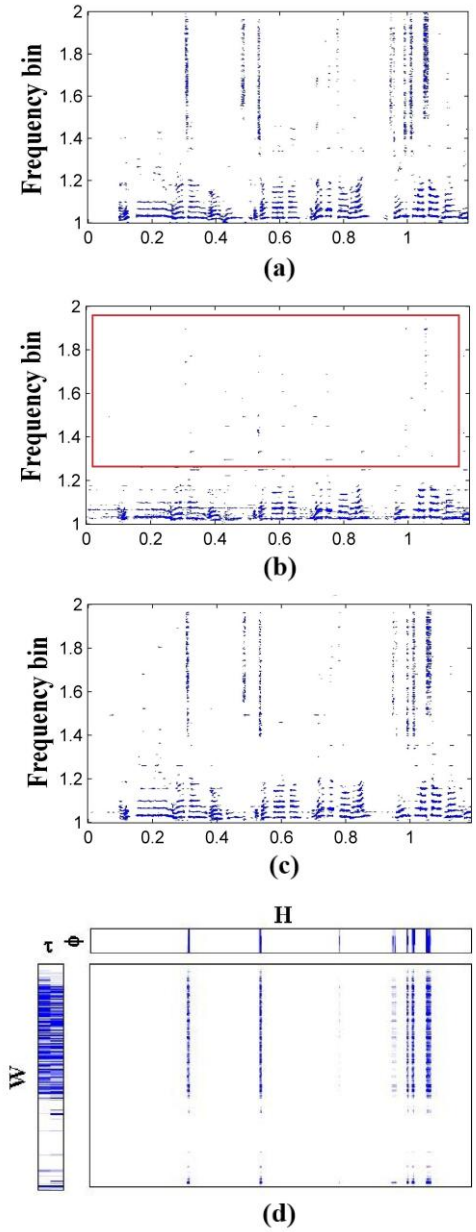

Fig. 7: Comparison between the spectrogram of the full-rank NMF, and the variable sparsity CNTF2D. (a) Spectrogram of the original source. (b) Spectrogram of the estimated source by using the full-rank NMF. (c) Spectrogram of the estimated source by using the variable sparsity CNTF2D. (d) One component of *W* and *H*, with their corresponding spectrogram for the variable sparsity CNTF2D.

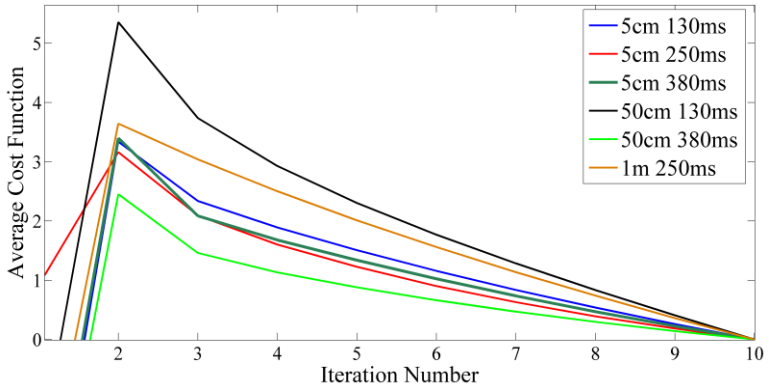

Fig. 8: Average cost function for different conditions.

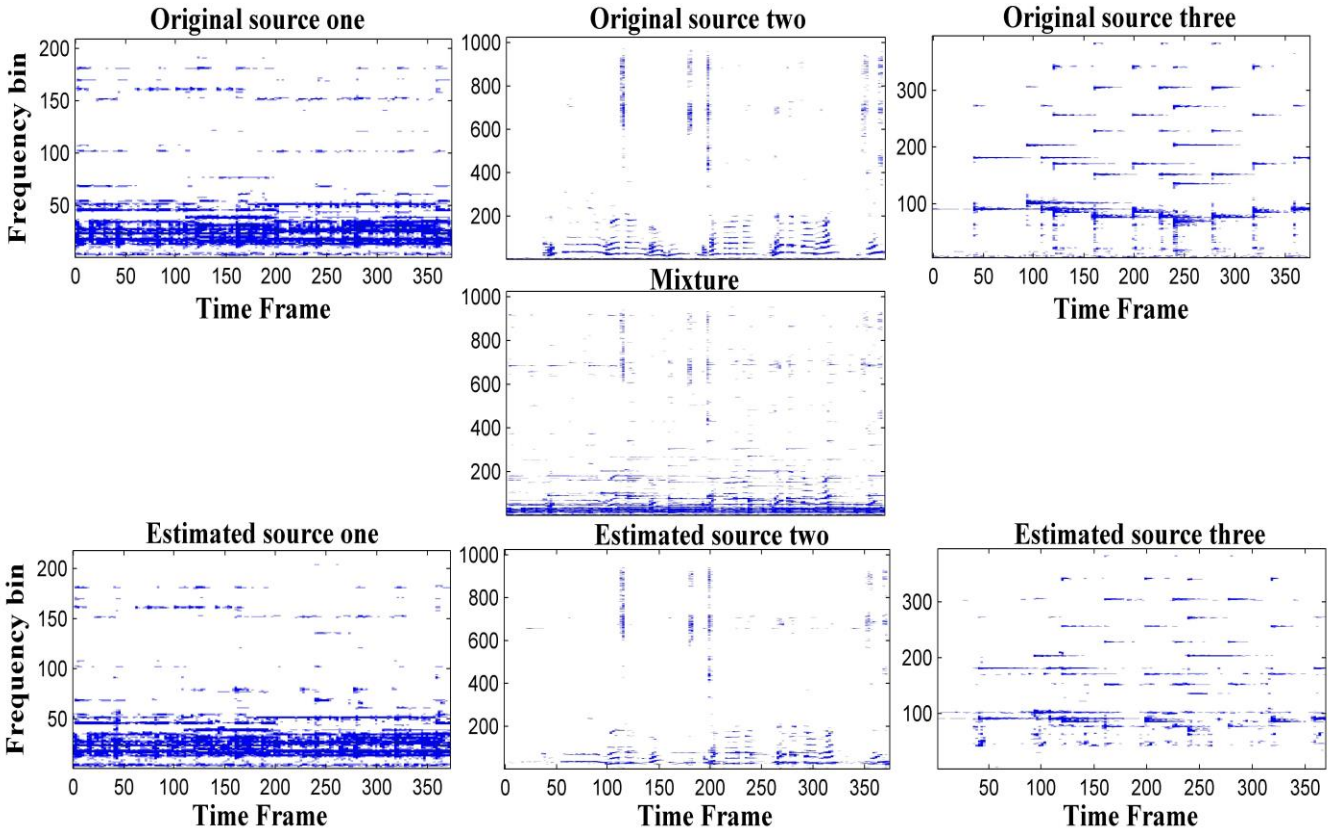

Fig. 9: Spectrogram of one of the mixtures and its original and estimated sources.

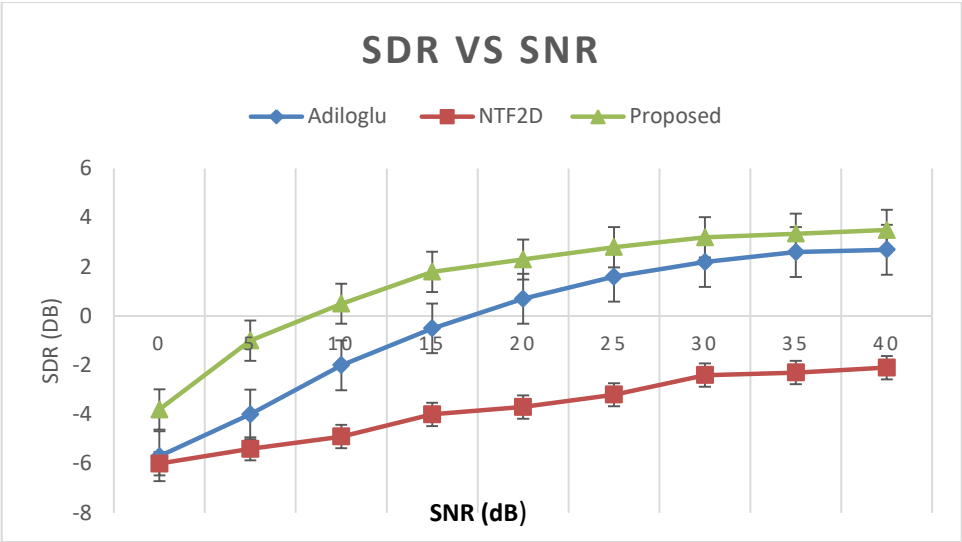

Fig. 10: Plot of SDR (dB) versus SNR (dB)